PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N 15/12, C07K 16/18, 14/47, C12Q 1/68 — A2

(11) International Publication Number: WO 99/38972

(43) International Publication Date: 5 August 1999 (05.08.99)

(21) International Application Number: PCT/US99/01619

(22) International Filing Date: 28 January 1999 (28.01.99)

(30) Priority Data:

60/072,910	28 January 1998 (28.01.98)	US
60/075,954	24 February 1998 (24.02.98)	US
60/080,114	31 March 1998 (31.03.98)	US
60/080,515	3 April 1998 (03.04.98)	US
60/080,666	3 April 1998 (03.04.98)	US

(71) Applicants *(for all designated States except US)*: CHIRON CORPORATION [US/US]; 4560 Horton Street, Emeryville, CA 94608 (US). HYSEQ INC. [US/US]; 675 Almanor Avenue, Sunnyvale, CA 94086 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: WILLIAMS, Lewis, T. [US/US]; 3 Miroflores, Tiburon, CA 94920 (US). ESCOBEDO, Jaime [CL/US]; 1470 Lavorna Road, Alamo, CA 94507 (US). INNIS, Michael, A. [US/US]; 315 Constance Place, Moraga, CA 94556 (US). GARCIA, Pablo, Dominguez [CL/US]; 882 Chenery Street, San Francisco, CA 94131 (US). SUDDUTH-KLINGER, Julie [US/US]; 280 Lexington Road, Kensington, CA 94707 (US). REINHARD, Christoph [DE/US]; 1633 Clinton Avenue, Alameda, CA 94501 (US). GIESE, Klause [DE/US]; 1009 Carolina Street, San Francisco, CA 94107 (US). RANDAZZO, Filippo [US/US]; 6363 Christie Avenue #2511, Emeryville, CA 94608 (US). KENNEDY, Giulia, C. [US/US]; 360 Castenada Avenue, San Francisco, CA 94116 (US). POT, David [CA/US]; 1565 5th Avenue #102, San Francisco, CA 94112 (US). KASSAM, Altaf [US/US]; 394 49th Street, Oakland, CA 94609 (US). LAMSON, George [US/US]; 232 Sandringham Drive, Moraga, CA 94556 (US). DRMANAC, Radoje [YU/US]; 850 East Greenwich Place, Palo Alto, CA 94303 (US). CRKVENJAKOV, Radomir [YU/US]; 762 Haverhill Drive, Sunnyvale, CA 94068 (US). DICKSON, Mark [US/US]; 1411 Gabilan Drive #B, Hollister, CA 95025 (US). DRMANAC, Snezana [YU/US]; 850 East Greenwich Place, Palo Alto, CA 94303 (US). LABAT, Ivan [YU/US]; 140 Acalanes Drive, Sunnyvale, CA 94086 (US). LESHKOWITZ, Dena [US/US]; 678 Durshire Way, Sunnyvale, CA 94087 (US). KITA, David [US/US]; 899 Bounty Drive, Foster City, CA 94404 (US). GARCIA, Veronica [ES/US]; 911 Shell Boulevard #102-0, Foster City, CA 96606 (US). JONES, William, Lee [US/US]; 4290 Albany Drive #P-146, San Jose, CA 95129 (US). STACHE-CRAIN, Birjit [DE/US]; 345 South Mary Avenue, Sunnyvale, CA 94086 (US).

(74) Agent: BLACKBURN, Robert, P.; Chiron Corporation, P.O. Box 8097, Emeryville, CA 94662-8097 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

Without international search report and to be republished upon receipt of that report.

(54) Title: HUMAN GENES AND GENE EXPRESSION PRODUCTS II

(57) Abstract

This invention relates to novel human polynucleotides and variants thereof, their encoded polypeptides and variants thereof, to genes corresponding to these polynucleotides and to proteins expressed by the genes. The invention also relates to diagnostic and therapeutic agents employing such novel human polynucleotides, their corresponding genes or gene products, e.g., these genes and proteins, including probes, antisense constructs, and antibodies.

HUMAN GENES AND GENE EXPRESSION PRODUCTS II

Field of the Invention

The present invention relates to novel polynucleotides, particularly to novel polynucleotides of human origin that are expressed in a selected cell type, are differentially expressed in one cell type relative to another cell type (*e.g.*, in cancerous cells, or in cells of a specific tissue origin) and/or share homology to polynucleotides encoding a gene product having an identified functional domain and/or activity.

Background of the Invention

Identification of novel polynucleotides, particularly those that encode an expressed gene product, is important in the advancement of drug discovery, diagnostic technologies, and the understanding of the progression and nature of complex diseases such as cancer. Identification of genes expressed in different cell types isolated from sources that differ in disease state or stage, developmental stage, exposure to various environmental factors, the tissue of origin, the species from which the tissue was isolated, and the like is key to identifying the genetic factors that are responsible for the phenotypes associated with these various differences

This invention provides novel human polynucleotides, the polypeptides encoded by these polynucleotides, and the genes and proteins corresponding to these novel polynucleotides.

Summary of the Invention

This invention relates to novel human polynucleotides and variants thereof, their encoded polypeptides and variants thereof, to genes corresponding to these polynucleotides and to proteins expressed by the genes. The invention also relates to diagnostic and therapeutic agents employing such novel human polynucleotides, their corresponding genes or gene products, *e.g.*, these genes and proteins, including probes, antisense constructs, and antibodies. The polynucleotides of the invention correspond to a polynucleotide comprising the sequence information of at least one of SEQ ID NOS: 1-3544, 3546-4510, 4512-4725, 4727-4748, and 4750-5252, which for convenience sake is referred to herein as "SEQ ID NOS:1-5252."

OCID: <WO__9938972A2_I_>

Accordingly, in one embodiment, the present invention features a library of polynucleotides, the library comprising the sequence information of at least one of "SEQ ID NOS:1-5252". In related aspects, the invention features a library provided on a nucleic acid array, or in a computer-readable format.

In one embodiment, the library is comprises a differentially expressed polynucleotide comprising a sequence selected from one of the differentially expressed polynucleotides disclosed herein. In specific related embodiments, the library comprises: 1) a polynucleotide that is differentially expressed in a human breast cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS:15, 36, 44, 45, 89, 146, 154, 159, 165, 172, 174, 183, 203, 261, 364, 366, 387, 419, 420, 496, 503, 510, 512, 529, 552, 560, 564, 570, 590, 606, 644, 646, 693, 707, 711, 726, 746, 754, 756, 875, 902, 921, 942, 990, 1095, 1104, 1122, 1131, 1142, 1170, 1184, 1205, 1286, 1289, 1354, 1387, 1435, 1535, 1751, 1764, 1777, 1795, 1860, 1869, 1882, 1890, 1915, 1933, 1934, 1979, 1980, 2007, 2023, 2040, 2059, 2223, 2245, 2300, 2325, 2409, 2462, 2488, 2486, and 2492; 2) a polynucleotide differentially expressed in a human colon cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS: , 33, 65, 228, 250, 252, 253, 280, 282, 355, 370, 387, 443, 460, 491, 545, 560, 581, 603, 680, 693, 703, 704, 716, 726, 746, 752, 753, 1095, 1104, 1205, 1241, 1264, 1354, 1387, 1401, 1442, 1514, 1734, 1742, 1780, 1851, 1899, 1915, 1954, 2024, 2066, 2262, and 2325; 3) a polynucleotide differentially expressed in a human lung cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS: 10, 54, 65, 171, 174, , 203, 252, 253, 254, , 285, 419, 420, 466, 491, 525, 526, 552, 571, 574, 590, 693, 700, 726, 742, 746, 861, 922, 990, 1088, 1288, 1355, 1417, 1422, 1444, 1454, 1570, 1597, 1979, 2007, 2024, 2034, 2038, 2126, and 2245; 4) a polynucleotide differentially expressed in growth factor-treated human microvascular endothelial cells (HMEC) relative to untreated HMEC, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS:648, 1899, and 648; or 5) polynucleotides that are differentially expressed across multiple libraries, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS: 65, 174, 203, 252, 253, 387, 419, 420, 491, 552, 560, 581, 590, 648, 693, 726, 746, 990, 1095, 1124, 1205, 1354, 1387, 1780, 1899, 1915, 1979, 2007, 2024, 2245, and 2325,

In another aspect, the invention features an isolated polynucleotide comprising a nucleotide sequence having at least 90% sequence identity to an identifying sequence of "SEQ ID NOS:1-5252" or a degenerate variant thereof. In related aspects, the invention features recombinant host cells and vectors comprising the polynucleotides of the invention, as well as isolated polypeptides encoded by the polynucleotides of the invention and antibodies that specifically bind such polypeptides.

In one embodiment, the invention features an isolated polynucleotide comprising a sequence encoding a polypeptide of a protein family or having a functional domain selected from the group consisting of: 4 transmembrane segments integral membrane proteins, 7 transmembrane receptors (rhodopsin family or secretin family), eukaryotic aspartyl proteases, ATPases associated with various cellular activities (AAA), Bcl-2, cyclins, DEAD box protein family, DEAD/H helicase protein family, MAP kinase kinase protein family, novel 3'5'-cyclic nucleotide phosphodiesterases, protein kinases, ras protein family, G-protein alpha subunit, phorbol esters/diacylglycerol binding proteins, protein kinase, trypsin, protein tyrosine phosphatase, wnt family of developmental signaling proteins, WW/rsp5/WWP domain containing proteins, Ank repeat, basic region plus leucine zipper domain, bromodomain, eukaryotic thiol (cysteine) protease active site, EF-hand, ETS domain, type II fibronectin collagen binding domain, thioredoxin, homeobox domain, TNFR/NGFR family cysteine-rich region, WD domain/G-beta repeats, zinc finger (C2H2 type), zinc finger (CCHC class), and zinc finger (C3HC4 type). In a specific related embodiment, the invention features a polynucleotide comprising a sequence of one of the SEQ ID NOS: listed in Table 3 or Table 20.

In another aspect, the invention features a method of detecting differentially expressed genes correlated with a cancerous state of a mammalian cell, where the method comprises the step of detecting at least one differentially expressed gene product in a test sample derived from a cell suspected of being cancerous, where the gene product is encoded by a gene corresponding to a sequence of at least one of the differentially expressed polynucleotides disclosed herein. Detection of the differentially expressed gene product is correlated with a cancerous state of the cell from which the test sample was derived. In one embodiment, the detecting is by hybridization of the test sample to a

XOCID: <WO___9938972A2_I_>

reference array, wherein the reference array comprises an identifying sequence of at least one of the differentially expressed polynucleotides disclosed herein.

In one embodiment of the method of the invention, the cell is a breast tissue derived cell, and the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:15, 36, 44, 45, 89, 146, 154, 159, 165, 172, 174, 183, 203, 261, 364, 366, 387, 419, 420, 496, 503, 510, 512, 529, 552, 560, 564, 570, 590, 606, 644, 646, 693, 707, 711, 726, 746, 754, 756, 875, 902, 921, 942, 990, 1095, 1104, 1122, 1131, 1142, 1170, 1184, 1205, 1286, 1289, 1354, 1387, 1435, 1535, 1751, 1764, 1777, 1795, 1860, 1869, 1882, 1890, 1915, 1933, 1934, 1979, 1980, 2007, 2023, 2040, 2059, 2223, 2245, 2300, 2325, 2409, 2462, 2486 2488, and 2492.

In another embodiment of the method of the invention, the cell is a colon tissue derived cell, and differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS: 65, 228, 252, 253, 280, 355, 491, 581, 603, 680, 693, 716, 726, 746, 752, 753, 1241, 1264, 1401, 1442, 1514, 1851, 1915, 2024, 2066, 33, 250, 282, 370, 387, 443, 460, 545, 560, 703, 704, 1095, 1104, 1205, 1354, 1387, 1734, 1742, 1780, 1899, 1954, 2262, and 2325.

In yet another embodiment of the method of the invention, the cell is a lung tissue derived cell, and differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:10, 54, 65, 171, 174, 203, 252, 253, 254, 285, 419, 420, 466, 491, 525, 526, 552, 571, 574, 590, 693, 700, 726, 742, 746, 861, 922, 990, 1088, 1288, 1355, 1417, 1422, 1444, 1454, 1570, 1597, 1979, 2007, 2024, 2034, 2038, 2126, and 2245.

In another embodiment, the cell is any of a lung, breast, or colon cell and the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:648 and 1899.

In still another embodiment, the cell is any of a breast, colon, or lung cell and the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS: 65, 174, 203, 252, 253, 387, 419, 420, 491, 552, 560, 581, 590, 648, 693, 726, 746, 990, 1095, 1124, 1205, 1354, 1387, , 1780, 1899, 1915, 1979, 2007, 2024, 2245, and 2325.

Other aspects and embodiments of the invention will be readily apparent to the ordinarily skilled artisan upon reading the description provided herein.

Detailed Description of the Invention

The invention relates to polynucleotides comprising the disclosed nucleotide sequences, to full length cDNA, mRNA and genes corresponding to these sequences, and to polypeptides and proteins encoded by these polynucleotides and genes.

Also included are polynucleotides that encode polypeptides and proteins encoded by the polynucleotides of the Sequence Listing. The various polynucleotides that can encode these polypeptides and proteins differ because of the degeneracy of the genetic code, in that most amino acids are encoded by more than one triplet codon. The identity of such codons is well-known in this art, and this information can be used for the construction of the polynucleotides within the scope of the invention.

Polynucleotides encoding polypeptides and proteins that are variants of the polypeptides and proteins encoded by the polynucleotides and related cDNA and genes are also within the scope of the invention. The variants differ from wild type protein in having one or more amino acid substitutions that either enhance, add, or diminish a biological activity of the wild type protein. Once the amino acid change is selected, a polynucleotide encoding that variant is constructed according to the invention.

The following detailed description describes the polynucleotide compositions encompassed by the invention, methods for obtaining cDNA or genomic DNA encoding a full-length gene product, expression of these polynucleotides and genes, identification of structural motifs of the polynucleotides and genes, identification of the function of a gene product encoded by a gene corresponding to a polynucleotide of the invention, use of the provided polynucleotides as probes and in mapping and in tissue profiling, use of the corresponding polypeptides and other gene products to raise antibodies, and use of the polynucleotides and their encoded gene products for therapeutic and diagnostic purposes.

I. Polynucleotide Compositions

The scope of the invention with respect to polynucleotide compositions includes, but is not necessarily limited to, polynucleotides having a sequence set forth in any one of

XCID: <WO___9938972A2_I_>

"SEQ ID NOS:1-5252"; polynucleotides obtained from the biological materials described herein or other biological sources (particularly human sources) by hybridization under stringent conditions (particularly conditions of high stringency); genes corresponding to the provided polynucleotides; variants of the provided polynucleotides and their corresponding genes, particularly those variants that retain a biological activity of the encoded gene product (*e.g.*, a biological activity ascribed to a gene product corresponding to the provided polynucleotides as a result of the assignment of the gene product to a protein family(ies) and/or identification of a functional domain present in the gene product). Other nucleic acid compositions contemplated by and within the scope of the present invention will be readily apparent to one of ordinary skill in the art when provided with the disclosure here.

The invention features polynucleotides that are expressed in cells of human tissue, specifically human colon, breast, and/or lung tissue. Novel nucleic acid compositions of the invention of particular interest comprise a sequence set forth in any one of "SEQ ID NOS:1-5252" or an identifying sequence thereof. An "identifying sequence" is a contiguous sequence of residues at least about 10 nt to about 20 nt in length, usually at least about 50 nt to about 100 nt in length, that uniquely identifies a polynucleotide sequence, *e.g.*, exhibits less than 90%, usually less than about 80% to about 85% sequence identity to any contiguous nucleotide sequence of more than about 20 nt. Thus, the subject novel nucleic acid compositions include full length cDNAs or mRNAs that encompass an identifying sequence of contiguous nucleotides from any one of "SEQ ID NOS:1-5252."

The polynucleotides of the invention also include polynucleotides having sequence similarity or sequence identity. Nucleic acids having sequence similarity are detected by hybridization under low stringency conditions, for example, at 50°C and 10XSSC (0.9 M saline/0.09 M sodium citrate) and remain bound when subjected to washing at 55°C in 1XSSC. Sequence identity can be determined by hybridization under stringent conditions, for example, at 50°C or higher and 0.1XSSC (9 mM saline/0.9 mM sodium citrate). Hybridization methods and conditions are well known in the art, see, *e.g.*, U.S. Patent No. 5,707,829. Nucleic acids that are substantially identical to the provided polynucleotide sequences, *e.g.* allelic variants, genetically altered versions of the gene, *etc.*, bind to the provided polynucleotide sequences ("SEQ ID NOS:1-5252") under stringent hybridization conditions. By using probes, particularly labeled probes of DNA sequences, one can

isolate homologous or related genes. The source of homologous genes can be any species, *e.g.* primate species, particularly human; rodents, such as rats and mice; canines, felines, bovines, ovines, equines, yeast, nematodes, *etc.*

Preferably, hybridization is performed using at least 15 contiguous nucleotides of at least one of "SEQ ID NOS:1-5252." That is, when at least 15 contiguous nucleotides of one of the disclosed SEQ ID NOs. is used as a probe, the probe will preferentially hybridize with a gene or mRNA (of the biological material) comprising the complementary sequence, allowing the identification and retrieval of the nucleic acids of the biological material that uniquely hybridize to the selected probe. Probes from more than one SEQ ID NO. will hybridize with the same gene or mRNA if the cDNA from which they were derived corresponds to one mRNA. Probes of more than 15 nucleotides can be used, but 15 nucleotides represents enough sequence for unique identification.

The polynucleotides of the invention also include naturally occurring variants of the nucleotide sequences (*e.g.*, degenerate variants, allelic variants, *etc.*). Variants of the polynucleotides of the invention are identified by hybridization of putative variants with nucleotide sequences disclosed herein, preferably by hybridization under stringent conditions For example, by using appropriate wash conditions, variants of the polynucleotides of the invention can be identified where the allelic variant exhibits at most about 25-30% base pair mismatches relative to the selected polynucleotide probe. In general, allelic variants contain 15-25% base pair mismatches, and can contain as little as even 5-15%, or 2-5%, or 1-2% base pair mismatches, as well as a single base-pair mismatch.

The invention also encompasses homologs corresponding to the polynucleotides of "SEQ ID NOS:1-5252", where the source of homologous genes can be any mammalian species, *e.g.*, primate species, particularly human; rodents, such as rats; canines, felines, bovines, ovines, equines, yeast, nematodes, etc. Between mammalian species, *e.g.*, human and mouse, homologs have substantial sequence similarity, *e.g.*, at least 75% sequence identity, usually at least 90%, more usually at least 95% between nucleotide sequences. Sequence similarity is calculated based on a reference sequence, which may be a subset of a larger sequence, such as a conserved motif, coding region, flanking region, *etc.* A reference sequence will usually be at least about 18 contiguous nt long, more usually at

DOCID: <WO__9938972A2_I_>

least about 30 nt long, and may extend to the complete sequence that is being compared. Algorithms for sequence analysis are known in the art, such as BLAST, described in Altschul *et al.*, *J. Mol. Biol.* (1990) *215*:403-10.

In general, variants of the invention have a sequence identity greater than at least about 65%, preferably at least about 75%, more preferably at least about 85%, and can be greater than at least about 90% or more as determined by the Smith-Waterman homology search algorithm as implemented in MPSRCH program (Oxford Molecular). For the purposes of this invention, a preferred method of calculating percent identity is the Smith-Waterman algorithm, using the following. Global DNA sequence identity must be greater than 65% as determined by the Smith-Waterman homology search algorithm as implemented in MPSRCH program (Oxford Molecular) using an affine gap search with the following search parameters: gap open penalty, 12; and gap extension penalty, 1.

The subject nucleic acids can be cDNAs or genomic DNAs, as well as fragments thereof, particularly fragments that encode a biologically active gene product and/or are useful in the methods disclosed herein (*e.g.*, in diagnosis, as a unique identifier of a differentially expressed gene of interest, *etc.*). The term "cDNA" as used herein is intended to include all nucleic acids that share the arrangement of sequence elements found in native mature mRNA species, where sequence elements are exons and 3' and 5' non-coding regions. Normally mRNA species have contiguous exons, with the intervening introns, when present, being removed by nuclear RNA splicing, to create a continuous open reading frame encoding a polypeptide of the invention.

A genomic sequence of interest comprises the nucleic acid present between the initiation codon and the stop codon, as defined in the listed sequences, including all of the introns that are normally present in a native chromosome. It can further include the 3' and 5' untranslated regions found in the mature mRNA. It can further include specific transcriptional and translational regulatory sequences, such as promoters, enhancers, *etc.*, including about 1 kb, but possibly more, of flanking genomic DNA at either the 5' and 3' end of the transcribed region. The genomic DNA can be isolated as a fragment of 100 kbp or smaller; and substantially free of flanking chromosomal sequence. The genomic DNA flanking the coding region, either 3' and 5', or internal regulatory sequences as sometimes

found in introns, contains sequences required for proper tissue, stage-specific, or disease-state specific expression.

The nucleic acid compositions of the subject invention can encode all or a part of the subject polypeptides. Double or single stranded fragments can be obtained from the DNA sequence by chemically synthesizing oligonucleotides in accordance with conventional methods, by restriction enzyme digestion, by PCR amplification, *etc.* Isolated polynucleotides and polynucleotide fragments of the invention comprise at least about 10, about 15, about 20, about 35, about 50, about 100, about 150 to about 200, about 250 to about 300, or about 350 contiguous nucleotides selected from the polynucleotide sequences as shown in "SEQ ID NOS:1-5252." For the most part, fragments will be of at least 15 nt, usually at least 18 nt or 25 nt, and up to at least about 50 contiguous nt in length or more. In a preferred embodiment, the polynucleotide molecules comprise a contiguous sequence of at least twelve nucleotides selected from the group consisting of the polynucleotides shown in "SEQ ID NOS:1-5252."

Probes specific to the polynucleotides of the invention can be generated using the polynucleotide sequences disclosed in "SEQ ID NOS:1-5252." The probes are preferably at least about 12, 15, 16, 18, 20, 22, 24, or 25 nucleotide fragment of a corresponding contiguous sequence of "SEQ ID NOS:1-5252", and can be less than 2, 1, 0.5, 0.1, or 0.05 kb in length. The probes can be synthesized chemically or can be generated from longer polynucleotides using restriction enzymes. The probes can be labeled, for example, with a radioactive, biotinylated, or fluorescent tag. Preferably, probes are designed based upon an identifying sequence of a polynucleotide of one of "SEQ ID NOS:1-5252." More preferably, probes are designed based on a contiguous sequence of one of the subject polynucleotides that remain unmasked following application of a masking program for masking low complexity (*e.g.*, XBLAST) to the sequence., *i.e.*, one would select an unmasked region, as indicated by the polynucleotides outside the poly-n stretches of the masked sequence produced by the masking program.

The polynucleotides of the subject invention are isolated and obtained in substantial purity, generally as other than an intact chromosome. Usually, the polynucleotides, either as DNA or RNA, will be obtained substantially free of other naturally-occurring nucleic acid sequences, generally being at least about 50%, usually at least about 90% pure and are

DCID: <WO___9938972A2_I_>

typically "recombinant", *e.g.*, flanked by one or more nucleotides with which it is not normally associated on a naturally occurring chromosome.

The polynucleotides of the invention can be provided as a linear molecule or within a circular molecule. They can be provided within autonomously replicating molecules (vectors) or within molecules without replication sequences. They can be regulated by their own or by other regulatory sequences, as is known in the art. The polynucleotides of the invention can be introduced into suitable host cells using a variety of techniques which are available in the art, such as transferrin polycation-mediated DNA transfer, transfection with naked or encapsulated nucleic acids, liposome-mediated DNA transfer, intracellular transportation of DNA-coated latex beads, protoplast fusion, viral infection, electroporation, gene gun, calcium phosphate-mediated transfection, and the like.

The subject nucleic acid compositions can be used to, for example, produce polypeptides, as probes for the detection of mRNA of the invention in biological samples (*e.g.*, extracts of human cells) to generate additional copies of the polynucleotides, to generate ribozymes or antisense oligonucleotides, and as single stranded DNA probes or as triple-strand forming oligonucleotides. The probes described herein can be used to, for example, determine the presence or absence of the polynucleotide sequences as shown in "SEQ ID NOS:1-5252" or variants thereof in a sample. These and other uses are described in more detail below.

Use of Polynucleotides to Obtain Full-Length cDNA and Full-Length Human Gene and Promoter Region

Full-length cDNA molecules comprising the disclosed polynucleotides are obtained as follows. A polynucleotide having a sequence of one of "SEQ ID NOS:1-5252", or a portion thereof comprising at least 12, 15, 18, or 20 nucleotides, is used as a hybridization probe to detect hybridizing members of a cDNA library using probe design methods, cloning methods, and clone selection techniques such as those described in U.S. Patent No. 5,654,173. Libraries of cDNA are made from selected tissues, such as normal or tumor tissue, or from tissues of a mammal treated with, for example, a pharmaceutical agent. Preferably, the tissue is the same as the tissue from which the polynucleotides of the invention were isolated, as both the polynucleotides described herein and the cDNA

represent expressed genes. Most preferably, the cDNA library is made from the biological material described herein in the Examples. Alternatively, many cDNA libraries are available commercially. (Sambrook *et al., Molecular Cloning: A Laboratory Manual, 2nd Ed.*, (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY). The choice of cell type for library construction can be made after the identity of the protein encoded by the gene corresponding to the polynucleotide of the invention is known. This will indicate which tissue and cell types are likely to express the related gene, and thus represent a suitable source for the mRNA for generating the cDNA. Where the provided polynucleotides are isolated from cDNA libraries, the libraries are prepared from mRNA of human colon cells, more preferably, human colon cancer cells, even more preferably, from a highly metastatic colon cell, Km12L4-A.

Techniques for producing and probing nucleic acid sequence libraries are described, for example, in Sambrook *et al., Molecular Cloning: A Laboratory Manual, 2nd Ed.*, (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. The cDNA can be prepared by using primers based on sequence from "SEQ ID NOS:1-5252." In one embodiment, the cDNA library can be made from only poly-adenylated mRNA. Thus, poly-T primers can be used to prepare cDNA from the mRNA.

Members of the library that are larger than the provided polynucleotides, and preferably that encompass the complete coding sequence of the native message, are obtained. In order to confirm that the entire cDNA has been obtained, RNA protection experiments are performed as follows. Hybridization of a full-length cDNA to an mRNA will protect the RNA from RNase degradation. If the cDNA is not full length, then the portions of the mRNA that are not hybridized will be subject to RNase degradation. This is assayed, as is known in the art, by changes in electrophoretic mobility on polyacrylamide gels, or by detection of released monoribonucleotides. Sambrook *et al., Molecular Cloning: A Laboratory Manual, 2nd Ed.*, (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. In order to obtain additional sequences 5' to the end of a partial cDNA, 5' RACE (*PCR Protocols: A Guide to Methods and Applications*, (1990) Academic Press, Inc.) is performed.

Genomic DNA is isolated using the provided polynucleotides in a manner similar to the isolation of full-length cDNAs. Briefly, the provided polynucleotides, or portions

DOCID: <WO___9938972A2_I_>

thereof, are used as probes to libraries of genomic DNA. Preferably, the library is obtained from the cell type that was used to generate the polynucleotides of the invention, but this is not essential. Most preferably, the genomic DNA is obtained from the biological material described herein in the Examples. Such libraries can be in vectors suitable for carrying large segments of a genome, such as P1 or YAC, as described in detail in Sambrook *et al.*, 9.4-9.30. In addition, genomic sequences can be isolated from human BAC libraries, which are commercially available from Research Genetics, Inc., Huntville, Alabama, USA, for example. In order to obtain additional 5' or 3' sequences, chromosome walking is performed, as described in Sambrook *et al.*, such that adjacent and overlapping fragments of genomic DNA are isolated. These are mapped and pieced together, as is known in the art, using restriction digestion enzymes and DNA ligase.

Using the polynucleotide sequences of the invention, corresponding full-length genes can be isolated using both classical and PCR methods to construct and probe cDNA libraries. Using either method, Northern blots, preferably, are performed on a number of cell types to determine which cell lines express the gene of interest at the highest level. Classical methods of constructing cDNA libraries are taught in Sambrook *et al.*, *supra*. With these methods, cDNA can be produced from mRNA and inserted into viral or expression vectors. Typically, libraries of mRNA comprising poly(A) tails can be produced with poly(T) primers. Similarly, cDNA libraries can be produced using the instant sequences as primers.

PCR methods are used to amplify the members of a cDNA library that comprise the desired insert. In this case, the desired insert will contain sequence from the full length cDNA that corresponds to the instant polynucleotides. Such PCR methods include gene trapping and RACE methods. Gene trapping entails inserting a member of a cDNA library into a vector. The vector then is denatured to produce single stranded molecules. Next, a substrate-bound probe, such a biotinylated oligo, is used to trap cDNA inserts of interest. Biotinylated probes can be linked to an avidin-bound solid substrate. PCR methods can be used to amplify the trapped cDNA. To trap sequences corresponding to the full length genes, the labeled probe sequence is based on the polynucleotide sequences of the invention. Random primers or primers specific to the library vector can be used to amplify the trapped cDNA. Such gene trapping techniques are described in Gruber *et al.*, WO

95/04745 and Gruber *et al.*, U.S. Pat. No. 5,500,356. Kits are commercially available to perform gene trapping experiments from, for example, Life Technologies, Gaithersburg, Maryland, USA.

"Rapid amplification of cDNA ends," or RACE, is a PCR method of amplifying cDNAs from a number of different RNAs. The cDNAs are ligated to an oligonucleotide linker, and amplified by PCR using two primers. One primer is based on sequence from the instant polynucleotides, for which full length sequence is desired, and a second primer comprises sequence that hybridizes to the oligonucleotide linker to amplify the cDNA. A description of this methods is reported in WO 97/19110. In preferred embodiments of RACE, a common primer is designed to anneal to an arbitrary adaptor sequence ligated to cDNA ends (Apte and Siebert, *Biotechniques* (1993) *15*:890-893; Edwards *et al.*, *Nuc. Acids Res.* (1991) *19*:5227-5232). When a single gene-specific RACE primer is paired with the common primer, preferential amplification of sequences between the single gene specific primer and the common primer occurs. Commercial cDNA pools modified for use in RACE are available.

Another PCR-based method generates full-length cDNA library with anchored ends without needing specific knowledge of the cDNA sequence. This method is described in WO 96/40998.

The promoter region of a gene generally is located 5' to the initiation site for RNA polymerase II. Hundreds of promoter regions contain the "TATA" box, a sequence such as TATTA or TATAA, which is sensitive to mutations. The promoter region can be obtained by performing 5' RACE using a primer from the coding region of the gene. Alternatively, the cDNA can be used as a probe for the genomic sequence, and the region 5' to the coding region is identified by "walking up." If the gene is highly expressed or differentially expressed, the promoter from the gene can be of use in a regulatory construct for a heterologous gene.

Once the full-length cDNA or gene is obtained, DNA encoding variants can be prepared by site-directed mutagenesis, described in detail in Sambrook *et al.*, 15.3-15.63. The choice of codon or nucleotide to be replaced can be based on disclosure herein on optional changes in amino acids to achieve altered protein structure and/or function.

DOCID: <WO___9938972A2_I_>

As an alternative method to obtaining DNA or RNA from a biological material, nucleic acid comprising nucleotides having the sequence of one or more polynucleotides of the invention can be synthesized. Thus, the invention encompasses nucleic acid molecules ranging in length from 15 nucleotides (corresponding to at least 15 contiguous nucleotides of one of "SEQ ID NOS:1-5252") up to a maximum length suitable for one or more biological manipulations, including replication and expression, of the nucleic acid molecule. The invention includes but is not limited to (a) nucleic acid having the size of a full gene, and comprising at least one of "SEQ ID NOS:1-5252;"; (b) the nucleic acid of (a) also comprising at least one additional gene, operably linked to permit expression of a fusion protein; (c) an expression vector comprising (a) or (b); (d) a plasmid comprising (a) or (b) ; and (e) a recombinant viral particle comprising (a) or (b). Once provided with the polynucleotides disclosed herein, construction or preparation of (a) - (e) are well within the skill in the art.

The sequence of a nucleic acid comprising at least 15 contiguous nucleotides of at least any one of "SEQ ID NOS:1-5252,", preferably the entire sequence of at least any one of "SEQ ID NOS:1-5252," is not limited and can be any sequence of A, T, G, and/or C (for DNA) and A, U, G, and/or C (for RNA) or modified bases thereof, including inosine and pseudouridine. The choice of sequence will depend on the desired function and can be dictated by coding regions desired, the intron-like regions desired, and the regulatory regions desired. Where the entire sequence of any one of "SEQ ID NOS:1-5252" is within the nucleic acid, the nucleic acid obtained is referred to herein as a polynucleotide comprising the sequence of any one of "SEQ ID NOS:1-5252."

II. Expression of Polypeptide Encoded by Full-Length cDNA or Full-Length Gene

The provided polynucleotide (*e.g.*, a polynucleotide having a sequence of one of "SEQ ID NOS:1-5252"), the corresponding cDNA, or the full-length gene is used to express a partial or complete gene product. Constructs of polynucleotides having sequences of "SEQ ID NOS:1-5252" can be generated synthetically. Alternatively, single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides is described by, *e.g.*, Stemmer *et al., Gene (Amsterdam)* (1995) *164(1)*:49-53. In this method, assembly PCR (the synthesis of long DNA sequences from

large numbers of oligodeoxyribonucleotides (oligos)) is described. The method is derived from DNA shuffling (Stemmer, *Nature* (1994) *370*:389-391), and does not rely on DNA ligase, but instead relies on DNA polymerase to build increasingly longer DNA fragments during the assembly process.

Appropriate polynucleotide constructs are purified using standard recombinant DNA techniques as described in, for example, Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual, 2nd Ed.*, (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY, and under current regulations described in United States Dept. of HHS, National Institute of Health (NIH) Guidelines for Recombinant DNA Research. The gene product encoded by a polynucleotide of the invention is expressed in any expression system, including, for example, bacterial, yeast, insect, amphibian and mammalian systems. Suitable vectors and host cells are described in U.S. Patent No. 5,654,173.

Bacteria. Expression systems in bacteria include those described in Chang *et al.*, *Nature* (1978) *275*:615; Goeddel *et al.*, *Nature* (1979) *281*:544; Goeddel *et al.*, *Nucleic Acids Res.* (1980) *8*:4057; EP 0 036,776; U.S. Patent No. 4,551,433; DeBoer *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1983) *80*:21-25; and Siebenlist *et al.*, *Cell* (1980) *20*:269.

Yeast. Expression systems in yeast include those described in Hinnen *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1978) *75*:1929; Ito *et al.*, *J. Bacteriol.* (1983) *153*:163; Kurtz *et al.*, *Mol. Cell. Biol.* (1986) *6*:142; Kunze *et al.*, *J. Basic Microbiol.* (1985) *25*:141; Gleeson *et al.*, *J. Gen. Microbiol.* (1986) *132*:3459; Roggenkamp *et al.*, *Mol. Gen. Genet.* (1986) *202*:302; Das *et al.*, *J. Bacteriol.* (1984) *158*:1165; De Louvencourt *et al.*, *J. Bacteriol.* (1983) *154*:737; Van den Berg *et al.*, *Bio/Technology* (1990) *8*:135; Kunze *et al.*, *J. Basic Microbiol.* (1985) *25*:141; Cregg *et al.*, *Mol. Cell. Biol.* (1985) *5*:3376; U.S. Patent Nos. 4,837,148 and 4,929,555; Beach and Nurse, *Nature* (1981) *300*:706; Davidow *et al.*, *Curr. Genet.* (1985) *10*:380; Gaillardin *et al.*, *Curr. Genet.* (1985) *10*:49; Ballance *et al.*, *Biochem. Biophys. Res. Commun.* (1983) *112*:284-289; Tilburn *et al.*, *Gene* (1983) *26*:205-221; Yelton *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1984) *81*:1470-1474; Kelly and Hynes, *EMBO J.* (1985) *4*:475479; EP 0 244,234; and WO 91/00357.

Insect Cells. Expression of heterologous genes in insects is accomplished as described in U.S. Patent No. 4,745,051; Friesen *et al.*, "The Regulation of Baculovirus Gene Expression", in: *The Molecular Biology Of Baculoviruses* (1986) (W. Doerfler, ed.);

DCID: <WO___9938972A2_I_>

EP 0 127,839; EP 0 155,476; and Vlak *et al.*, *J. Gen. Virol.* (1988) *69*:765-776; Miller *et al.*, *Ann. Rev. Microbiol.* (1988) *42*:177; Carbonell *et al.*, *Gene* (1988) *73*:409; Maeda *et al.*, *Nature* (1985) *315*:592-594; Lebacq-Verheyden *et al.*, *Mol. Cell. Biol.* (1988) *8*:3129; Smith *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1985) *82*:8844; Miyajima *et al.*, *Gene* (1987) *58*:273; and Martin *et al.*, *DNA* (1988) *7*:99. Numerous baculoviral strains and variants and corresponding permissive insect host cells from hosts are described in Luckow *et al.*, *Bio/Technology* (1988) *6*:47-55, Miller *et al.*, *Generic Engineering* (1986) *8*:277-279, and Maeda *et al.*, *Nature* (1985) *315*:592-594.

Mammalian Cells. Mammalian expression is accomplished as described in Dijkema *et al.*, *EMBO J.* (1985) *4*:761, Gorman *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1982) *79*:6777, Boshart *et al.*, *Cell* (1985) *41*:521 and U.S. Patent No. 4,399,216. Other features of mammalian expression are facilitated as described in Ham and Wallace, *Meth. Enz.* (1979) *58*:44, Barnes and Sato, *Anal. Biochem.* (1980) *102*:255, U.S. Patent Nos. 4,767,704, 4,657,866, 4,927,762, 4,560,655, WO 90/103430, WO 87/00195, and U.S. RE 30,985.

Polynucleotide molecules comprising a polynucleotide sequence provided herein propagated by placing the molecule in a vector. Viral and non-viral vectors are used, including plasmids. The choice of plasmid will depend on the type of cell in which propagation is desired and the purpose of propagation. Certain vectors are useful for amplifying and making large amounts of the desired DNA sequence. Other vectors are suitable for expression in cells in culture. Still other vectors are suitable for transfer and expression in cells in a whole animal or person. The choice of appropriate vector is well within the skill of the art. Many such vectors are available commercially. The partial or full-length polynucleotide is inserted into a vector typically by means of DNA ligase attachment to a cleaved restriction enzyme site in the vector. Alternatively, the desired nucleotide sequence can be inserted by homologous recombination in vivo. Typically this is accomplished by attaching regions of homology to the vector on the flanks of the desired nucleotide sequence. Regions of homology are added by ligation of oligonucleotides, or by polymerase chain reaction using primers comprising both the region of homology and a portion of the desired nucleotide sequence, for example.

The polynucleotides set forth in "SEQ ID NOS:1-5252" or their corresponding full-length polynucleotides are linked to regulatory sequences as appropriate to obtain the desired expression properties. These can include promoters (attached either at the 5' end of the sense strand or at the 3' end of the antisense strand), enhancers, terminators, operators, repressors, and inducers. The promoters can be regulated or constitutive. In some situations it may be desirable to use conditionally active promoters, such as tissue-specific or developmental stage-specific promoters. These are linked to the desired nucleotide sequence using the techniques described above for linkage to vectors. Any techniques known in the art can be used.

When any of the above host cells, or other appropriate host cells or organisms, are used to replicate and/or express the polynucleotides or nucleic acids of the invention, the resulting replicated nucleic acid, RNA, expressed protein or polypeptide, is within the scope of the invention as a product of the host cell or organism. The product is recovered by any appropriate means known in the art.

Once the gene corresponding to a selected polynucleotide is identified, its expression can be regulated in the cell to which the gene is native. For example, an endogenous gene of a cell can be regulated by an exogenous regulatory sequence as disclosed in U.S. Patent No. 5,641,670.

III. Identification of Functional and Structural Motifs of Novel Genes

A. Screening Polynucleotide Sequences and Amino Acid Sequences Against Publicly Available Databases

Translations of the nucleotide sequence of the provided polynucleotides, cDNAs or full genes can be aligned with individual known sequences. Similarity with individual sequences can be used to determine the activity of the polypeptides encoded by the polynucleotides of the invention. For example, sequences that show similarity with a chemokine sequence can exhibit chemokine activities. Also, sequences exhibiting similarity with more than one individual sequence can exhibit activities that are characteristic of either or both individual sequences.

The full length sequences and fragments of the polynucleotide sequences of the nearest neighbors can be used as probes and primers to identify and isolate the full length

DOCID: <WO___9938972A2_I_>

sequence corresponding to provided polynucleotides. The nearest neighbors can indicate a tissue or cell type to be used to construct a library for the full-length sequences corresponding to the provided polynucleotides..

Typically, a selected polynucleotide is translated in all six frames to determine the best alignment with the individual sequences. The sequences disclosed herein in the Sequence Listing are in a 5' to 3' orientation and translation in three frames can be sufficient (with a few specific exceptions as described in the Examples). These amino acid sequences are referred to, generally, as query sequences, which will be aligned with the individual sequences. Databases with individual sequences are described in "Computer Methods for Macromolecular Sequence Analysis" *Methods in Enzymology* (1996) *266*, Doolittle, Academic Press, Inc., a division of Harcourt Brace & Co., San Diego, California, USA. Databases include Genbank, EMBL, and DNA Database of Japan (DDBJ).

Query and individual sequences can be aligned using the methods and computer programs described above, and include BLAST, available over the world wide web at http://ww.ncbi.nlm.nih.gov/BLAST/. Another alignment algorithm is Fasta, available in the Genetics Computing Group (GCG) package, Madison, Wisconsin, USA, a wholly owned subsidiary of Oxford Molecular Group, Inc. Other techniques for alignment are described in Doolittle, *supra*. Preferably, an alignment program that permits gaps in the sequence is utilized to align the sequences. The Smith-Waterman is one type of algorithm that permits gaps in sequence alignments. See *Meth. Mol. Biol.* (1997) *70*: 173-187. Also, the GAP program using the Needleman and Wunsch alignment method can be utilized to align sequences. An alternative search strategy uses MPSRCH software, which runs on a MASPAR computer. MPSRCH uses a Smith-Waterman algorithm to score sequences on a massively parallel computer. This approach improves ability to identify sequences that are distantly related matches, and is especially tolerant of small gaps and nucleotide sequence errors. Amino acid sequences encoded by the provided polynucleotides can be used to search both protein and DNA databases.

Results of individual and query sequence alignments can be divided into three categories, high similarity, weak similarity, and no similarity. Individual alignment results ranging from high similarity to weak similarity provide a basis for determining polypeptide activity and/or structure. Parameters for categorizing individual results include: percentage

of the alignment region length where the strongest alignment is found, percent sequence identity, and p value.

The percentage of the alignment region length is calculated by counting the number of residues of the individual sequence found in the region of strongest alignment, *e.g.*, contiguous region of the individual sequence that contains the greatest number of residues that are identical to the residues of the corresponding region of the aligned query sequence. This number is divided by the total residue length of the query sequence to calculate a percentage. For example, a query sequence of 20 amino acid residues might be aligned with a 20 amino acid region of an individual sequence. The individual sequence might be identical to amino acid residues 5, 9-15, and 17-19 of the query sequence. The region of strongest alignment is thus the region stretching from residue 9-19, an 11 amino acid stretch. The percentage of the alignment region length is: 11 (length of the region of strongest alignment) divided by (query sequence length) 20 or 55%.

Percent sequence identity is calculated by counting the number of amino acid matches between the query and individual sequence and dividing total number of matches by the number of residues of the individual sequences found in the region of strongest alignment. Thus, the percent identity in the example above would be 10 matches divided by 11 amino acids, or approximately, 90.9%

P value is the probability that the alignment was produced by chance. For a single alignment, the p value can be calculated according to Karlin *et al.*, *Proc. Natl. Acad. Sci.* (1990) *87*:2264 and Karlin *et al.*, *Proc. Natl. Acad. Sci.* (1993) *90*. The p value of multiple alignments using the same query sequence can be calculated using an heuristic approach described in Altschul *et al.*, *Nat. Genet.* (1994) *6*:119. Alignment programs such as BLAST program can calculate the p value.

Another factor to consider for determining identity or similarity is the location of the similarity or identity. Strong local alignment can indicate similarity even if the length of alignment is short. Sequence identity scattered throughout the length of the query sequence also can indicate a similarity between the query and profile sequences. The boundaries of the region where the sequences align can be determined according to Doolittle, *supra*; BLAST or FAST programs; or by determining the area where sequence identity is highest.

DOCID: <WO___9938972A2_I_>

High Similarity. In general, in alignment results considered to be of high similarity, the percent of the alignment region length is typically at least about 55% of total length query sequence; more typically, at least about 58%; even more typically; at least about 60% of the total residue length of the query sequence. Usually, percent length of the alignment region can be as much as about 62%; more usually, as much as about 64%; even more usually, as much as about 66%. Further, for high similarity, the region of alignment, typically, exhibits at least about 75% of sequence identity; more typically, at least about 78%; even more typically; at least about 80% sequence identity. Usually, percent sequence identity can be as much as about 82%; more usually, as much as about 84%; even more usually, as much as about 86%.

The p value is used in conjunction with these methods. If high similarity is found, the query sequence is considered to have high similarity with a profile sequence when the p value is less than or equal to about 10^{-2}; more usually; less than or equal to about 10^{-3}; even more usually; less than or equal to about 10^{-4}. More typically, the p value is no more than about 10^{-5}; more typically; no more than or equal to about 10^{-10}; even more typically; no more than or equal to about 10^{-15} for the query sequence to be considered high similarity.

Weak Similarity. In general, where alignment results considered to be of weak similarity, there is no minimum percent length of the alignment region nor minimum length of alignment. A better showing of weak similarity is considered when the region of alignment is, typically, at least about 15 amino acid residues in length; more typically, at least about 20; even more typically; at least about 25 amino acid residues in length. Usually, length of the alignment region can be as much as about 30 amino acid residues; more usually, as much as about 40; even more usually, as much as about 60 amino acid residues. Further, for weak similarity, the region of alignment, typically, exhibits at least about 35% of sequence identity; more typically, at least about 40%; even more typically; at least about 45% sequence identity. Usually, percent sequence identity can be as much as about 50%; more usually, as much as about 55%; even more usually, as much as about 60%.

If low similarity is found, the query sequence is considered to have weak similarity with a profile sequence when the p value is usually less than or equal to about 10^{-2}; more usually; less than or equal to about 10^{-3}; even more usually; less than or equal to about 10^{-4}. More

typically, the p value is no more than about 10^{-5}; more usually; no more than or equal to about 10^{-10}; even more usually; no more than or equal to about 10^{-15} for the query sequence to be considered weak similarity.

Similarity Determined by Sequence Identity Alone. Sequence identity alone can be used to determine similarity of a query sequence to an individual sequence and can indicate the activity of the sequence. Such an alignment, preferably, permits gaps to align sequences. Typically, the query sequence is related to the profile sequence if the sequence identity over the entire query sequence is at least about 15%; more typically, at least about 20%; even more typically, at least about 25%; even more typically, at least about 50%. Sequence identity alone as a measure of similarity is most useful when the query sequence is usually, at least 80 residues in length; more usually, 90 residues; even more usually, at least 95 amino acid residues in length. More typically, similarity can be concluded based on sequence identity alone when the query sequence is preferably 100 residues in length; more preferably, 120 residues in length; even more preferably, 150 amino acid residues in length.

Determining Activity from Alignments with Profile and Multiple Aligned Sequences. Translations of the provided polynucleotides can be aligned with amino acid profiles that define either protein families or common motifs. Also, translations of the provided polynucleotides can be aligned to multiple sequence alignments (MSA) comprising the polypeptide sequences of members of protein families or motifs. Similarity or identity with profile sequences or MSAs can be used to determine the activity of the gene products (*e.g.*, polypeptides) encoded by the provided polynucleotides or corresponding cDNA or genes. For example, sequences that show an identity or similarity with a chemokine profile or MSA can exhibit chemokine activities.

Profiles can designed manually by (1) creating an MSA, which is an alignment of the amino acid sequence of members that belong to the family and (2) constructing a statistical representation of the alignment. Such methods are described, for example, in Birney *et al.*, *Nucl. Acid Res.* (1996) *24(14)*: 2730-2739. MSAs of some protein families and motifs are publicly available. For example, http://genome.wustl.edu/Pfam/ includes MSAs of 547 different families and motifs. These MSAs are described also in Sonnhammer *et al.*, *Proteins* (1997) *28*: 405-420. Other sources over the world wide web

OCID: <WO__9938972A2_I_>

include the site at http://www.embl-heidelberg.de/argos/ali/ali.html; alternatively, a message can be sent to ALI@EMBL-HEIDELBERG.DE for the information. A brief description of these MSAs is reported in Pascarella *et al.*, *Prot. Eng.* (1996) *9(3)*:249-251. Techniques for building profiles from MSAs are described in Sonnhammer *et al.*, *supra*; Birney *et al.*, *supra*; and "Computer Methods for Macromolecular Sequence Analysis," *Methods in Enzymology* (1996) *266*, Doolittle, Academic Press, Inc., a division of Harcourt Brace & Co., San Diego, California, USA.

Similarity between a query sequence and a protein family or motif can be determined by (a) comparing the query sequence against the profile and/or (b) aligning the query sequence with the members of the family or motif. Typically, a program such as Searchwise is used to compare the query sequence to the statistical representation of the multiple alignment, also known as a profile. The program is described in Birney *et al.*, *supra*. Other techniques to compare the sequence and profile are described in Sonnhammer *et al.*, *supra* and Doolittle, *supra*.

Next, methods described by Feng *et al.*, *J. Mol. Evol.* (1987) *25*:351 and Higgins *et al.*, *CABIOS* (1989) *5*:151 can be used align the query sequence with the members of a family or motif, also known as a MSA. Computer programs, such as PILEUP, can be used. See Feng *et al.*, *infra*. In general, the following factors are used to determine if a similarity between a query sequence and a profile or MSA exists: (1) number of conserved residues found in the query sequence, (2) percentage of conserved residues found in the query sequence, (3) number of frameshifts, and (4) spacing between conserved residues.

Some alignment programs that both translate and align sequences can make any number of frameshifts when translating the nucleotide sequence to produce the best alignment. The fewer frameshifts needed to produce an alignment, the stronger the similarity or identity between the query and profile or MSAs. For example, a weak similarity resulting from no frameshifts can be a better indication of activity or structure of a query sequence, than a strong similarity resulting from two frameshifts. Preferably, three or fewer frameshifts are found in an alignment; more preferably two or fewer frameshifts; even more preferably, one or fewer frameshifts; even more preferably, no frameshifts are found in an alignment of query and profile or MSAs.

Conserved residues are those amino acids found at a particular position in all or some of the family or motif members. For example, most chemokines contain four conserved cysteines. Alternatively, a position is considered conserved if only a certain class of amino acids is found in a particular position in all or some of the family members. For example, the N-terminal position can contain a positively charged amino acid, such as lysine, arginine, or histidine.

Typically, a residue of a polypeptide is conserved when a class of amino acids or a single amino acid is found at a particular position in at least about 40% of all class members; more typically, at least about 50%; even more typically, at least about 60% of the members. Usually, a residue is conserved when a class or single amino acid is found in at least about 70% of the members of a family or motif; more usually, at least about 80%; even more usually, at least about 90%; even more usually, at least about 95%.

A residue is considered conserved when three unrelated amino acids are found at a particular position in the some or all of the members; more usually, two unrelated amino acids. These residues are conserved when the unrelated amino acids are found at particular positions in at least about 40% of all class member; more typically, at least about 50%; even more typically, at least about 60% of the members. Usually, a residue is conserved when a class or single amino acid is found in at least about 70% of the members of a family or motif; more usually, at least about 80%; even more usually, at least about 90%; even more usually, at least about 95%.

A query sequence has similarity to a profile or MSA when the query sequence comprises at least about 25% of the conserved residues of the profile or MSA; more usually, at least about 30%; even more usually; at least about 40%. Typically, the query sequence has a stronger similarity to a profile sequence or MSA when the query sequence comprises at least about 45% of the conserved residues of the profile or MSA; more typically, at least about 50%; even more typically; at least about 55%.

B. Screening Polynucleotide and Amino Acid Sequences Against Protein Profiles

The identify and function of the gene that correlates to a polynucleotide described herein can be determined by screening the polynucleotides or their corresponding amino acid sequences against profiles of protein families. Such profiles focus on common

)OCID: <WO__9938972A2_I_>

structural motifs among proteins of each family. Publicly available profiles are described above in Section IVA. Additional or alternative profiles are described below.

In comparing a novel polynucleotide with known sequences, several alignment tools are available. Examples include PileUp, which creates a multiple sequence alignment, and is described in Feng *et al., J. Mol. Evol.* (1987) *25*:351. Another method, GAP, uses the alignment method of Needleman *et al., J. Mol. Biol.* (1970) *48*:443. GAP is best suited for global alignment of sequences. A third method, BestFit, functions by inserting gaps to maximize the number of matches using the local homology algorithm of Smith *et al., Adv. Appl. Math.* (1981) *2*:482.

C. Identification of Secreted & Membrane-Bound Polypeptides

Both secreted and membrane-bound polypeptides of the present invention are of particular interest. For example, levels of secreted polypeptides can be assayed in body fluids that are convenient, such as blood, urine, prostatic fluid and semen. Membrane-bound polypeptides are useful for constructing vaccine antigens or inducing an immune response. Such antigens would comprise all or part of the extracellular region of the membrane-bound polypeptides. Because both secreted and membrane-bound polypeptides comprise a fragment of contiguous hydrophobic amino acids, hydrophobicity predicting algorithms can be used to identify such polypeptides.

A signal sequence is usually encoded by both secreted and membrane-bound polypeptide genes to direct a polypeptide to the surface of the cell. The signal sequence usually comprises a stretch of hydrophobic residues. Such signal sequences can fold into helical structures. Membrane-bound polypeptides typically comprise at least one transmembrane region that possesses a stretch of hydrophobic amino acids that can transverse the membrane. Some transmembrane regions also exhibit a helical structure. Hydrophobic fragments within a polypeptide can be identified by using computer algorithms. Such algorithms include Hopp & Woods, *Proc. Natl. Acad. Sci. USA* (1981) *78*:3824-3828; Kyte & Doolittle, *J. Mol. Biol.* (1982) *157*: 105-132; and RAOAR algorithm, Degli Esposti *et al., Eur. J. Biochem.* (1990) *190*: 207-219.

Another method of identifying secreted and membrane-bound polypeptides is to translate the polynucleotides of the invention in all six frames and determine if at least 8

contiguous hydrophobic amino acids are present. Those translated polypeptides with at least 8; more typically, 10; even more typically, 12 contiguous hydrophobic amino acids are considered to be either a putative secreted or membrane bound polypeptide. Hydrophobic amino acids include alanine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, threonine, tryptophan, tyrosine, and valine.

IV. Identification of the Function of an Expression Product of a Full-Length Gene Corresponding to a Polynucleotide

Ribozymes, antisense constructs, and dominant negative mutants can be used to determine function of the expression product of a gene corresponding to a polynucleotide provided herein. These methods and compositions are particularly useful where the provided novel polynucleotide exhibits no significant or substantial homology to a sequence encoding a gene of known function. Antisense molecules and ribozymes can be constructed from synthetic polynucleotides. Typically, the phosphoramidite method of oligonucleotide synthesis is used. See Beaucage *et al.*, *Tet. Lett.* (1981) *22*:1859 and U.S. Patent No. 4,668,777. Automated devices for synthesis are available to create oligonucleotides using this chemistry. Examples of such devices include Biosearch 8600, Models 392 and 394 by Applied Biosystems, a division of Perkin-Elmer Corp., Foster City, California, USA; and Expedite by Perceptive Biosystems, Framingham, Massachusetts, USA. Synthetic RNA, phosphate analog oligonucleotides, and chemically derivatized oligonucleotides can also be produced, and can be covalently attached to other molecules. RNA oligonucleotides can be synthesized, for example, using RNA phosphoramidites. This method can be performed on an automated synthesizer, such as Applied Biosystems, Models 392 and 394, Foster City, California, USA. See Applied Biosystems User Bulletin 53 and Ogilvie *et al.*, *Pure & Applied Chem.* (1987) *59*:325.

Phosphorothioate oligonucleotides can also be synthesized for antisense construction. A sulfurizing reagent, such as tetraethylthiruam disulfide (TETD) in acetonitrile can be used to convert the internucleotide cyanoethyl phosphite to the phosphorothioate triester within 15 minutes at room temperature. TETD replaces the iodine reagent, while all other reagents used for standard phosphoramidite chemistry

DCID: <WO___9938972A2_I_>

remain the same. Such a synthesis method can be automated using Models 392 and 394 by Applied Biosystems, for example.

Oligonucleotides of up to 200 nucleotides can be synthesized, more typically, 100 nucleotides, more typically 50 nucleotides; even more typically 30 to 40 nucleotides. These synthetic fragments can be annealed and ligated together to construct larger fragments. See, for example, Sambrook *et al.*, *supra*.

A. Ribozymes

Trans-cleaving catalytic RNAs (ribozymes) are RNA molecules possessing endoribonuclease activity. Ribozymes are specifically designed for a particular target, and the target message must contain a specific nucleotide sequence. They are engineered to cleave any RNA species site-specifically in the background of cellular RNA. The cleavage event renders the mRNA unstable and prevents protein expression. Importantly, ribozymes can be used to inhibit expression of a gene of unknown function for the purpose of determining its function in an in vitro or in vivo context, by detecting the phenotypic effect.

One commonly used ribozyme motif is the hammerhead, for which the substrate sequence requirements are minimal. Design of the hammerhead ribozyme is disclosed in Usman *et al., Current Opin. Struct. Biol.* (1996) *6*:527. Ribozymes can also be prepared and used as described in Long *et al., FASEB J.* (1993) *7*:25; Symons, *Ann. Rev. Biochem.* (1992) *61*:641; Perrotta *et al., Biochem.* (1992) *31*:16; Ojwang *et al., Proc. Natl. Acad. Sci. (USA)* (1992) *89*:10802; and U.S. Patent No. 5,254,678. Ribozyme cleavage of HIV-I RNA is described in U.S. Patent No. 5,144,019; methods of cleaving RNA using ribozymes is described in U.S. Patent No. 5,116,742; and methods for increasing the specificity of ribozymes are described in U.S. Patent No. 5,225,337 and Koizumi *et al., Nucleic Acid Res.* (1989) *17*:7059. Preparation and use of ribozyme fragments in a hammerhead structure are also described by Koizumi *et al., Nucleic Acids Res.* (1989) *17*:7059. Preparation and use of ribozyme fragments in a hairpin structure are described by Chowrira and Burke, *Nucleic Acids Res.* (1992) *20*:2835. Ribozymes can also be made by rolling transcription as described in Daubendiek and Kool, *Nat. Biotechnol.* (1997) *15(3)*:273.

The hybridizing region of the ribozyme can be modified or can be prepared as a branched structure as described in Horn and Urdea, *Nucleic Acids Res.* (1989) *17*:6959. The basic structure of the ribozymes can also be chemically altered in ways familiar to those skilled in the art, and chemically synthesized ribozymes can be administered as synthetic oligonucleotide derivatives modified by monomeric units. In a therapeutic context, liposome mediated delivery of ribozymes improves cellular uptake, as described in Birikh *et al., Eur. J. Biochem.* (1997) *245*:1.

Using the polynucleotide sequences of the invention and methods known in the art, ribozymes are designed to specifically bind and cut the corresponding mRNA species. Ribozymes thus provide a means to inhibit the expression of any of the proteins encoded by the disclosed polynucleotides or their full-length genes. The full-length gene need not be known in order to design and use specific inhibitory ribozymes. In the case of a polynucleotide or full-length cDNA of unknown function, ribozymes corresponding to that nucleotide sequence can be tested in vitro for efficacy in cleaving the target transcript. Those ribozymes that effect cleavage in vitro are further tested in vivo. The ribozyme can also be used to generate an animal model for a disease, as described in Birikh *et al., supra*. An effective ribozyme is used to determine the function of the gene of interest by blocking its transcription and detecting a change in the cell. Where the gene is found to be a mediator in a disease, an effective ribozyme is designed and delivered in a gene therapy for blocking transcription and expression of the gene.

Therapeutic and functional genomic applications of ribozymes proceed beginning with knowledge of a portion of the coding sequence of the gene to be inhibited. Thus, for many genes, a partial polynucleotide sequence provides adequate sequence for constructing an effective ribozyme. A target cleavage site is selected in the target sequence, and a ribozyme is constructed based on the 5' and 3' nucleotide sequences that flank the cleavage site. Retroviral vectors are engineered to express monomeric and multimeric hammerhead ribozymes targeting the mRNA of the target coding sequence. These monomeric and multimeric ribozymes are tested in vitro for an ability to cleave the target mRNA. A cell line is stably transduced with the retroviral vectors expressing the ribozymes, and the transduction is confirmed by Northern blot analysis and reverse-transcription polymerase chain reaction (RT-PCR). The cells are screened for inactivation of the target mRNA by

DCID: <WO__9938972A2_I_>

such indicators as reduction of expression of disease markers or reduction of the gene product of the target mRNA.

B. Antisense

Antisense nucleic acids are designed to specifically bind to RNA, resulting in the formation of RNA-DNA or RNA-RNA hybrids, with an arrest of DNA replication, reverse transcription or messenger RNA translation. Antisense polynucleotides based on a selected polynucleotide sequence can interfere with expression of the corresponding gene. Antisense polynucleotides are typically generated within the cell by expression from antisense constructs that contain the antisense strand as the transcribed strand. Antisense polynucleotides based on the disclosed polynucleotides will bind and/or interfere with the translation of mRNA comprising a sequence complementary to the antisense polynucleotide. The expression products of control cells and cells treated with the antisense construct are compared to detect the protein product of the gene corresponding to the polynucleotide upon which the antisense construct is based. The protein is isolated and identified using routine biochemical methods.

Given the extensive background literature and clinical experience in antisense therapy, one skilled in the art can use selected polynucleotides of the invention as additional potential therapeutics. The choice of polynucleotide can be narrowed by first testing them for binding to "hot spot" regions of the genome of cancerous cells. If a polynucleotide is identified as binding to a "hot spot", testing the polynucleotide as an antisense compound in the corresponding cancer cells clearly is warranted.

C. Dominant Negative Mutations

As an alternative method for identifying function of the gene corresponding to a polynucleotide disclosed herein, dominant negative mutations are readily generated for corresponding proteins that are active as homomultimers. A mutant polypeptide will interact with wild-type polypeptides (made from the other allele) and form a non-functional multimer. Thus, a mutation is in a substrate-binding domain, a catalytic domain, or a cellular localization domain. Preferably, the mutant polypeptide will be overproduced. Point mutations are made that have such an effect. In addition, fusion of different polypeptides of various lengths to the terminus of a protein can yield dominant negative mutants. General strategies are available for making dominant negative mutants (see, *e.g.*,

Herskowitz, *Nature* (1987) *329*:219). Such techniques can be used to create loss of function mutations, which are useful for determining protein function.

V. Construction of Polypeptides of the Invention and Variants Thereof

The polypeptides of the invention include those encoded by the disclosed polynucleotides. These polypeptides can also be encoded by nucleic acids that, by virtue of the degeneracy of the genetic code, are not identical in sequence to the disclosed polynucleotides. Thus, the invention includes within its scope a polypeptide encoded by a polynucleotide having the sequence of any one of "SEQ ID NOS:1-5252" or a variant thereof.

In general, the term "polypeptide" as used herein refers to both the full length polypeptide encoded by the recited polynucleotide, the polypeptide encoded by the gene represented by the recited polynucleotide, as well as portions or fragments thereof. "Polypeptides" also includes variants of the naturally occurring proteins, where such variants are homologous or substantially similar to the naturally occurring protein, and can be of an origin of the same or different species as the naturally occurring protein (*e.g.*, human, murine, or some other species that naturally expresses the recited polypeptide, usually a mammalian species). In general, variant polypeptides have a sequence that has at least about 80%, usually at least about 90%, and more usually at least about 98% sequence identity with a differentially expressed polypeptide of the invention, as measured by BLAST using the parameters described above. The variant polypeptides can be naturally or non-naturally glycosylated, *i.e.*, the polypeptide has a glycosylation pattern that differs from the glycosylation pattern found in the corresponding naturally occurring protein.

The invention also encompasses homologs of the disclosed polypeptides (or fragments thereof) where the homologs are isolated from other species, *i.e.* other animal or plant species, where such homologs, usually mammalian species, *e.g.* rodents, such as mice, rats; domestic animals, *e.g.*, horse, cow, dog, cat; and humans. By homolog is meant a polypeptide having at least about 35%, usually at least about 40% and more usually at least about 60% amino acid sequence identity a particular differentially expressed protein as identified above, where sequence identity is determined using the BLAST algorithm, with the parameters described *supra*.

XOCID: <WO__9938972A2_I_>

In general, the polypeptides of the subject invention are provided in a non-naturally occurring environment, *e.g.* are separated from their naturally occurring environment. In certain embodiments, the subject protein is present in a composition that is enriched for the protein as compared to a control. As such, purified polypeptide is provided, where by purified is meant that the protein is present in a composition that is substantially free of non-differentially expressed polypeptides, where by substantially free is meant that less than 90%, usually less than 60% and more usually less than 50% of the composition is made up of non-differentially expressed polypeptides.

Also within the scope of the invention are variants; variants of polypeptides include mutants, fragments, and fusions. Mutants can include amino acid substitutions, additions or deletions. The amino acid substitutions can be conservative amino acid substitutions or substitutions to eliminate non-essential amino acids, such as to alter a glycosylation site, a phosphorylation site or an acetylation site, or to minimize misfolding by substitution or deletion of one or more cysteine residues that are not necessary for function. Conservative amino acid substitutions are those that preserve the general charge, hydrophobicity/hydrophilicity, and/or steric bulk of the amino acid substituted. For example, substitutions between the following groups are conservative: Gly/Ala, Val/Ile/Leu, Asp/Glu, Lys/Arg, Asn/Gln, Ser/Cys, Thr, and Phe/Trp/Tyr.

Variants can be designed so as to retain biological activity of a particular region of the protein (*e.g.*, a functional domain and/or, where the polypeptide is a member of a protein family, a region associated with a consensus sequence). In a non-limiting example, Osawa *et al.*, *Biochem. Mol. Int.* (1994) *34*:1003, discusses the actin binding region of a protein from several different species. The actin binding regions of the these species are considered homologous based on the fact that they have amino acids that fall within "homologous residue groups." Homologous residues are judged according to the following groups (using single letter amino acid designations): STAG; ILVMF; HRK; DEQN; and FYW. For example, and S, a T, an A or a G can be in a position and the function (in this case actin binding) is retained.

Additional guidance on amino acid substitution is available from studies of protein evolution. Go *et al*, *Int. J. Peptide Protein Res.* (1980) *15*:211, classified amino acid residue sites as interior or exterior depending on their accessibility. More frequent

substitution on exterior sites was confirmed to be general in eight sets of homologous protein families regardless of their biological functions and the presence or absence of a prosthetic group. Virtually all types of amino acid residues had higher mutabilities on the exterior than in the interior. No correlation between mutability and polarity was observed of amino acid residues in the interior and exterior, respectively. Amino acid residues were classified into one of three groups depending on their polarity: polar (Arg, Lys, His, Gln, Asn, Asp, and Glu); weak polar (Ala, Pro, Gly, Thr, and Ser), and nonpolar (Cys, Val, Met, Ile, Leu, Phe, Tyr, and Trp). Amino acid replacements during protein evolution were very conservative: 88% and 76% of them in the interior or exterior, respectively, were within the same group of the three. Inter-group replacements are such that weak polar residues are replaced more often by nonpolar residues in the interior and more often by polar residues on the exterior.

Additional guidance for production of polypeptide variants is provided in Querol *et al.*, *Prot. Eng.* (1996) *9*:265, which provides general rules for amino acid substitutions to enhance protein thermostability. New glycosylation sites can be introduced as discussed in Olsen and Thomsen, *J. Gen. Microbiol.* (1991) *137*:579. An additional disulfide bridge can be introduced, as discussed by Perry and Wetzel, *Science* (1984) *226*:555; Pantoliano *et al.*, *Biochemistry* (1987) *26*:2077; Matsumura *et al.*, *Nature* (1989) *342*:291; Nishikawa *et al.*, *Protein Eng.* (1990) *3*:443; Takagi *et al.*, *J. Biol. Chem.* (1990) *265*:6874; Clarke *et al.*, *Biochemistry* (1993) *32*:4322; and Wakarchuk *et al.*, *Protein Eng.* (1994) *7*:1379. Metal binding sites can be introduced, according to Toma *et al.*, *Biochemistry* (1991) *30*:97, and Haezerbrouck *et al.*, *Protein Eng.* (1993) *6*:643. Substitutions with prolines in loops can be made according to Masul *et al.*, *Appl. Env. Microbiol.* (1994) *60*:3579; and Hardy *et al.*, *FEBS Lett. 317*:89.

Cysteine-depleted muteins are considered variants within the scope of the invention. These variants can be constructed according to methods disclosed in U.S. Patent No. 4,959,314, which discloses substitution of cysteines with other amino acids, and methods for assaying biological activity and effect of the substitution. Such methods are suitable for proteins according to this invention that have cysteine residues suitable for such substitutions, for example to eliminate disulfide bond formation.

IOCID: <WO__9938972A2_I_>

Variants also include fragments of the polypeptides disclosed herein, particularly biologically active fragments and/or fragments corresponding to functional domains. Fragments of interest will typically be at least about 10 aa to at least about 15 aa in length, usually at least about 50 aa in length, and can be as long as 300 aa in length or longer, but will usually not exceed about 1000 aa in length, where the fragment will have a stretch of amino acids that is identical to a polypeptide encoded by a polynucleotide having a sequence of any "SEQ ID NOS:1-5252", or a homolog thereof.

The protein variants described herein are encoded by polynucleotides that are within the scope of the invention. The genetic code can be used to select the appropriate codons to construct the corresponding variants.

VI. Computer-Related Embodiments

In general, a library of polynucleotides is a collection of sequence information, which information is provided in either biochemical form (*e.g.*, as a collection of polynucleotide molecules), or in electronic form (*e.g.*, as a collection of polynucleotide sequences stored in a computer-readable form, as in a computer system and/or as part of a computer program). The sequence information of the polynucleotides can be used in a variety of ways, *e.g.*, as a resource for gene discovery, as a representation of sequences expressed in a selected cell type (*e.g.*, cell type markers), and/or as markers of a given disease or disease state. In general, a disease marker is a representation of a gene product that is present in all cells affected by disease either at an increased or decreased level relative to a normal cell (*e.g.*, a cell of the same or similar type that is not substantially affected by disease). For example, a polynucleotide sequence in a library can be a polynucleotide that represents an mRNA, polypeptide, or other gene product encoded by the polynucleotide, that is either overexpressed or underexpressed in a breast ductal cell affected by cancer relative to a normal (*i.e.*, substantially disease-free) breast cell.

The nucleotide sequence information of the library can be embodied in any suitable form, *e.g.*, electronic or biochemical forms. For example, a library of sequence information embodied in electronic form includes an accessible computer data file (or, in biochemical form, a collection of nucleic acid molecules) that contains the representative nucleotide sequences of genes that are differentially expressed (*e.g.*, overexpressed or underexpressed)

as between, for example, i) a cancerous cell and a normal cell; ii) a cancerous cell and a dysplastic cell; iii) a cancerous cell and a cell affected by a disease or condition other than cancer; iv) a metastatic cancerous cell and a normal cell and/or non-metastatic cancerous cell; v) a malignant cancerous cell and a non-malignant cancerous cell (or a normal cell) and/or vi) a dysplastic cell relative to a normal cell. Other combinations and comparisons of cells affected by various diseases or stages of disease will be readily apparent to the ordinarily skilled artisan. Biochemical embodiments of the library include a collection of nucleic acids that have the sequences of the genes in the library, where the nucleic acids can correspond to the entire gene in the library or to a fragment thereof, as described in greater detail below.

The polynucleotide libraries of the subject invention include sequence information of a plurality of polynucleotide sequences, where at least one of the polynucleotides has a sequence of any of "SEQ ID NOS:1-5252." By plurality is meant at least 2, usually at least 3 and can include up to all of "SEQ ID NOS:1-5252." The length and number of polynucleotides in the library will vary with the nature of the library, *e.g.*, if the library is an oligonucleotide array, a cDNA array, a computer database of the sequence information, etc.

Where the library is an electronic library, the nucleic acid sequence information can be present in a variety of media. "Media" refers to a manufacture, other than an isolated nucleic acid molecule, that contains the sequence information of the present invention. Such a manufacture provides the genome sequence or a subset thereof in a form that can be examined by means not directly applicable to the sequence as it exists in a nucleic acid. For example, the nucleotide sequence of the present invention, *e.g.* the nucleic acid sequences of any of the polynucleotides of "SEQ ID NOS:1-5252," can be recorded on computer readable media, *e.g.* any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as a floppy disc, a hard disc storage medium, and a magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. One of skill in the art can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising a recording of the present sequence information.

DOCID: <WO___9938972A2_I_>

"Recorded" refers to a process for storing information on computer readable medium, using any such methods as known in the art. Any convenient data storage structure can be chosen, based on the means used to access the stored information. A variety of data processor programs and formats can be used for storage, *e.g.* word processing text file, database format, *etc.* In addition to the sequence information, electronic versions of the libraries of the invention can be provided in conjunction or connection with other computer-readable information and/or other types of computer-readable files (*e.g.*, searchable files, executable files, *etc*, including, but not limited to, for example, search program software, *etc.*.).

By providing the nucleotide sequence in computer readable form, the information can be accessed for a variety of purposes. Computer software to access sequence information is publicly available. For example, the BLAST (Altschul *et al.*, *supra.*) and BLAZE (Brutlag *et al. Comp. Chem.* (1993) 17:203) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the genome that contain homology to ORFs from other organisms.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention. The data storage means can comprise any manufacture comprising a recording of the present sequence information as described above, or a memory access means that can access such a manufacture.

"Search means" refers to one or more programs implemented on the computer-based system, to compare a target sequence or target structural motif with the stored sequence information. Search means are used to identify fragments or regions of the genome that match a particular target sequence or target motif. A variety of known algorithms are publicly known and commercially available, *e.g.* MacPattern (EMBL), BLASTN and BLASTX (NCBI). A "target sequence" can be any DNA or amino acid

sequence of six or more nucleotides or two or more amino acids, preferably from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues.

A "target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration that is formed upon the folding of the target motif, or on consensus sequences of regulatory or active sites. There are a variety of target motifs known in the art. Protein target motifs include, but arc not limited to, enzyme active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, hairpin structures, promoter sequences and other expression elements such as binding sites for transcription factors.

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. One format for an output means ranks fragments of the genome possessing varying degrees of homology to a target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences and identifies the degree of sequence similarity contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer based systems of the present invention.

As discussed above, the "library" of the invention also encompasses biochemical libraries of the polynucleotides of "SEQ ID NOS:1-5252," *e.g.*, collections of nucleic acids representing the provided polynucleotides. The biochemical libraries can take a variety of forms, *e.g.*, a solution of cDNAs, a pattern of probe nucleic acids stably associated with a surface of a solid support (*i.e.*, an array) and the like. Of particular interest are nucleic acid arrays in which one or more of "SEQ ID NOS:1-5252"is represented on the array. By array is meant a an article of manufacture that has at least a substrate with at least two distinct nucleic acid targets on one of its surfaces, where the number of distinct nucleic acids can be considerably higher, typically being at least 10 nt, usually at least 20 nt and often at least 25 nt. A variety of different array formats have been developed and are known to those of

OCID: <WO__9938972A2_I_>

skill in the art, including those described in 5,242,974; 5,384,261; 5,405,783; 5,412,087; 5,424,186; 5,429,807; 5,436,327; 5,445,934; 5,472,672; 5,527,681; 5,529,756; 5,545,531; 5,554,501; 5,556,752; 5,561,071; 5,599,895; 5,624,711; 5,639,603; 5,658,734; WO 93/17126; WO 95/11995; WO 95/35505; EP 742287; and EP 799897. The arrays of the subject invention find use in a variety of applications, including gene expression analysis, drug screening, mutation analysis and the like, as disclosed in the above-listed exemplary patent documents.

In addition to the above nucleic acid libraries, analogous libraries of polypeptides are also provided, where the where the polypeptides of the library will represent at least a portion of the polypeptides encoded by "SEQ ID NOS:1-5252."

VII. Utilities

A. Use of Polynucleotide Probes in Mapping, and in Tissue Profiling

Polynucleotide probes, generally comprising at least 12 contiguous nucleotides of a polynucleotide as shown in the Sequence Listing, are used for a variety of purposes, such as chromosome mapping of the polynucleotide and detection of transcription levels. Additional disclosure about preferred regions of the disclosed polynucleotide sequences is found in the Examples. A probe that hybridizes specifically to a polynucleotide disclosed herein should provide a detection signal at least 5-, 10-, or 20-fold higher than the background hybridization provided with other unrelated sequences.

Probes in Detection of Expression Levels. Nucleotide probes are used to detect expression of a gene corresponding to the provided polynucleotide. In Northern blots, mRNA is separated electrophoretically and contacted with a probe. A probe is detected as hybridizing to an mRNA species of a particular size. The amount of hybridization is quantitated to determine relative amounts of expression, for example under a particular condition. Probes are used for in situ hybridization to cells to detect expression. Probes can also be used *in vivo* for diagnostic detection of hybridizing sequences. Probes are typically labeled with a radioactive isotope. Other types of detectable labels can be used such as chromophores, fluors, and enzymes. Other examples of nucleotide hybridization assays are described in WO92/02526 and U.S. Patent No. 5,124,246.

The Polymerase Chain Reaction (PCR) is another means for detecting small amounts of target nucleic acids (see, *e.g.*, Mullis *et al.*, *Meth. Enzymol.* (1987) *155*:335; U.S. Patent No. 4,683,195; and U.S. Patent No. 4,683,202). Two primer polynucleotides nucleotides hybridize with the target nucleic acids and are used to prime the reaction. The primers can be composed of sequence within or 3' and 5' to the polynucleotides of the Sequence Listing. Alternatively, if the primers are 3' and 5' to these polynucleotides, they need not hybridize to them or the complements. A thermostable polymerase creates copies of target nucleic acids from the primers using the original target nucleic acids as a template. After a large amount of target nucleic acids is generated by the polymerase, it is detected by methods such as Southern blots. When using the Southern blot method, the labeled probe will hybridize to a polynucleotide of the Sequence Listing or complement.

Furthermore, mRNA or cDNA can be detected by traditional blotting techniques described in Sambrook *et al.*, "Molecular Cloning: A Laboratory Manual" (New York, Cold Spring Harbor Laboratory, 1989). mRNA or cDNA generated from mRNA using a polymerase enzyme can be purified and separated using gel electrophoresis. The nucleic acids on the gel are then blotted onto a solid support, such as nitrocellulose. The solid support is exposed to a labeled probe and then washed to remove any unhybridized probe. Next, the duplexes containing the labeled probe are detected. Typically, the probe is labeled with radioactivity.

Mapping. Polynucleotides of the present invention are used to identify a chromosome on which the corresponding gene resides. Such mapping can be useful in identifying the function of the polynucleotide-related gene by its proximity to other genes with known function. Function can also be assigned to the polynucleotide-related gene when particular syndromes or diseases map to the same chromosome. For example, use of polynucleotide probes in identification and quantification of nucleic acid sequence aberrations is described in U.S. Patent No. 5,783,387.

For example, fluorescence in situ hybridization (FISH) on normal metaphase spreads facilitates comparative genomic hybridization to allow total genome assessment of changes in relative copy number of DNA sequences. See Schwartz and Samad, *Curr. Opin. Biotechnol.* (1994) *8*:70; Kallioniemi *et al.*, *Sem. Cancer Biol.* (1993) *4*:41; Valdes

ICID: <WO___9938972A2_I_>

et al., *Methods in Molecular Biology* (1997) *68*:1, Boultwood, ed., Human Press, Totowa, NJ.

Polynucleotides are mapped to particular chromosomes using, for example, radiation hybrids or chromosome-specific hybrid panels. See Leach *et al.*, *Advances in Genetics*, (1995) *33*:63-99; Walter *et al.*, *Nature Genetics* (1994) *7*:22; Walter and Goodfellow, *Trends in Genetics* (1992) *9*:352. Panels for radiation hybrid mapping are available from Research Genetics, Inc., Huntsville, Alabama, USA. Databases for markers using various panels are available via the world wide web at http:/F/shgc-www.stanford.edu; and http://www-genome.wi.mit.edu/cgi-bin/contig/rhmapper.pl. The statistical program RHMAP can be used to construct a map based on the data from radiation hybridization with a measure of the relative likelihood of one order versus another. RHMAP is available via the world wide web at http://www.sph.umich.edu/group/statgen/software.

In addition, commercial programs are available for identifying regions of chromosomes commonly associated with disease, such as cancer. Polynucleotides based on the polynucleotides of the invention can be used to probe these regions. For example, if through profile searching a provided polynucleotide is identified as corresponding to a gene encoding a kinase, its ability to bind to a cancer-related chromosomal region will suggest its role as a kinase in one or more stages of tumor cell development/growth. Although some experimentation would be required to elucidate the role, the polynucleotide constitutes a new material for isolating a specific protein that has potential for developing a cancer diagnostic or therapeutic.

Tissue Typing or Profiling. Expression of specific mRNA corresponding to the provided polynucleotides can vary in different cell types and can be tissue-specific. This variation of mRNA levels in different cell types can be exploited with nucleic acid probe assays to determine tissue types. For example, PCR, branched DNA probe assays, or blotting techniques utilizing nucleic acid probes substantially identical or complementary to polynucleotides listed in the Sequence Listing can determine the presence or absence of the corresponding cDNA or mRNA.

For example, a metastatic lesion is identified by its developmental organ or tissue source by identifying the expression of a particular marker of that organ or tissue. If a

polynucleotide is expressed only in a specific tissue type, and a metastatic lesion is found to express that polynucleotide, then the developmental source of the lesion has been identified. Expression of a particular polynucleotide is assayed by detection of either the corresponding mRNA or the protein product. Immunological methods, such as antibody staining, are used to detect a particular protein product. Hybridization methods can be used to detect particular mRNA species, including but not limited to in situ hybridization and Northern blotting.

Use of Polymorphisms. A polynucleotide of the invention will be useful in forensics, genetic analysis, mapping, and diagnostic applications if the corresponding region of a gene is polymorphic in the human population. Particular polymorphic forms of the provided polynucleotides can be used to either identify a sample as deriving from a suspect or rule out the possibility that the sample derives from the suspect. Any means for detecting a polymorphism in a gene are used, including but not limited to electrophoresis of protein polymorphic variants, differential sensitivity to restriction enzyme cleavage, and hybridization to allele-specific probes.

B. Antibody Production

Expression products of a polynucleotide of the invention, the corresponding mRNA or cDNA, or the corresponding complete gene are prepared and used for raising antibodies for experimental, diagnostic, and therapeutic purposes. For polynucleotides to which a corresponding gene has not been assigned, this provides an additional method of identifying the corresponding gene. The polynucleotide or related cDNA is expressed as described above, and antibodies are prepared. These antibodies are specific to an epitope on the polypeptide encoded by the polynucleotide, and can precipitate or bind to the corresponding native protein in a cell or tissue preparation or in a cell-free extract of an in vitro expression system.

Immunogens for raising antibodies are prepared by mixing the polypeptides encoded by the polynucleotides of the present invention with adjuvants. Alternatively, polypeptides are made as fusion proteins to larger immunogenic proteins. Polypeptides are also covalently linked to other larger immunogenic proteins, such as keyhole limpet hemocyanin. Immunogens are typically administered intradermally, subcutaneously, or intramuscularly. Immunogens are administered to experimental animals such as rabbits,

XOCID: <WO___9938972A2_I_>

sheep, and mice, to generate antibodies. Optionally, the animal spleen cells are isolated and fused with myeloma cells to form hybridomas which secrete monoclonal antibodies. Such methods are well known in the art. According to another method known in the art, the selected polynucleotide is administered directly, such as by intramuscular injection, and expressed in vivo. The expressed protein generates a variety of protein-specific immune responses, including production of antibodies, comparable to administration of the protein.

Preparations of polyclonal and monoclonal antibodies specific for polypeptides encoded by a selected polynucleotide are made using standard methods known in the art. The antibodies specifically bind to epitopes present in the polypeptides encoded by polynucleotides disclosed in the Sequence Listing. Typically, at least 6, 8, 10, or 12 contiguous amino acids are required to form an epitope. However, epitopes which involve non-contiguous amino acids may require more, for example at least 15, 25, or 50 amino acids. A short sequence of a polynucleotide may then be unsuitable for use as an epitope to raise antibodies for identifying the corresponding novel protein, because of the potential for cross-reactivity with a known protein. However, the antibodies can be useful for other purposes, particularly if they identify common structural features of a known protein and a novel polypeptide encoded by a polynucleotide of the invention.

Antibodies that specifically bind to human polypeptides encoded by the provided polypeptides should provide a detection signal at least 5-, 10-, or 20-fold higher than a detection signal provided with other proteins when used in Western blots or other immunochemical assays. Preferably, antibodies that specifically polypeptides of the invention do not bind to other proteins in immunochemical assays at detectable levels and can immunoprecipitate the specific polypeptide from solution.

To test for the presence of serum antibodies to the polypeptide of the invention in a human population, human antibodies are purified by methods well known in the art. Preferably, the antibodies are affinity purified by passing antiserum over a column to which the corresponding selected polypeptide or fusion protein is bound. The bound antibodies can then be eluted from the column, for example using a buffer with a high salt concentration.

In addition to the antibodies discussed above, genetically engineered antibody derivatives are made, such as single chain antibodies, according to methods well known in the art.

C. Use of Polynucleotides to Construct Arrays for Diagnostics

Polynucleotide arrays provide a high throughput technique that can assay a large number of polynucleotide sequences in a sample. This technology can be used as a diagnostic and as a tool to test for differential expression to determine function of an encoded protein. Arrays can be created by spotting polynucleotide probes onto a substrate (*e.g.*, glass, nitrocelllose, *etc.*) in a two-dimensional matrix or array having bound probes. The probes can be bound to the substrate by either covalent bonds or by non-specific interactions, such as hydrophobic interactions. Samples of polynucleotides can be detectably labeled (*e.g.*, using radioactive or fluorescent labels) and then hybridized to the probes. Double stranded polynucleotides, comprising the labeled sample polynucleotides bound to probe polynucleotides, can be detected once the unbound portion of the sample is washed away. Techniques for constructing arrays and methods of using these arrays are described in EP No. 0 799 897; PCT No. WO 97/29212; PCT No. WO 97/27317; EP No. 0 785 280; PCT No. WO 97/02357; U.S. Pat. No. 5,593,839; U.S. Pat. No. 5,578,832; EP No. 0 728 520; U.S. Pat. No. 5,599,695; EP No. 0 721 016; U.S. Pat. No. 5,556,752; PCT No. WO 95/22058; and U.S. Pat. No. 5,631,734.

As discussed in some detail above, arrays can be used to examine differential expression of genes and can be used to determine gene function. For example, arrays of the instant polynucleotide sequences can be used to determine if any of the provided polynucleotides are differentially expressed between a test cell and control cell (*e.g.*, cancer cells and normal cells). For example, high expression of a particular message in a cancer cell, which is not observed in a corresponding normal cell, can indicate a cancer specific protein. Exemplary uses of arrays are further described in, for example, Pappalarado *et al.*, *Sem. Radiation Oncol.* (1998) *8*:217; and Ramsay *Nature Biotechnol.* (1998) *16*:40.

D. Differential Expression

The polynucleotides of the invention can also be used to detect differences in expression levels between two cells, *e.g.*, as a method to identify abnormal or diseased tissue in a human. For polynucleotides corresponding to profiles of protein families, the

)OCID: <WO__9938972A2_I_>

choice of tissue can be selected according to the putative biological function. In general, the expression of a gene corresponding to a specific polynucleotide is compared between a first tissue that is suspected of being diseased and a second, normal tissue of the human. The tissue suspected of being abnormal or diseased can be derived from a different tissue type of the human, but preferably it is derived from the same tissue type; for example an intestinal polyp or other abnormal growth should be compared with normal intestinal tissue. The normal tissue can be the same tissue as that of the test sample, or any normal tissue of the patient, especially those that express the polynucleotide-related gene of interest (*e.g.*, brain, thymus, testis, heart, prostate, placenta, spleen, small intestine, skeletal muscle, pancreas, and the mucosal lining of the colon). A difference between the polynucleotide-related gene, mRNA, or protein in the two tissues which are compared, for example in molecular weight, amino acid or nucleotide sequence, or relative abundance, indicates a change in the gene, or a gene which regulates it, in the tissue of the human that was suspected of being diseased. Examples of detection of differential expression and its use in diagnosis of cancer are described in U.S. Patent Nos. 5,688,641 and 5,677,125.

The polynucleotide-related genes in the two tissues are compared by any means known in the art. For example, the two genes can be sequenced, and the sequence of the gene in the tissue suspected of being diseased compared with the gene sequence in the normal tissue. The genes corresponding to a provided polynucleotide, or portions thereof, in the two tissues are amplified, for example using nucleotide primers based on the nucleotide sequence shown in the Sequence Listing, using the polymerase chain reaction. The amplified genes or portions of genes are hybridized to detectably labeled nucleotide probes selected from a nucleotide sequence shown in the Sequence Listing. A difference in the nucleotide sequence of the isolated gene in the tissue suspected of being diseased compared with the normal nucleotide sequence suggests a role of the gene product encoded by the subject polynucleotide in the disease, and provides guidance for preparing a therapeutic agent.

Alternatively, mRNA corresponding to a provided polynucleotide in the two tissues is compared. PolyA$^+$RNA is isolated from the two tissues as is known in the art. For example, one of skill in the art can readily determine differences in the size or amount of mRNA transcripts between the two tissues using Northern blots and detectably labeled

nucleotide probes selected from the nucleotide sequence shown in the Sequence Listing. Increased or decreased expression of a given mRNA in a tissue sample suspected of being diseased, compared with the expression of the same mRNA in a normal tissue, suggests that the expressed protein has a role in the disease, and also provides a lead for preparing a therapeutic agent.

The comparison can also be accomplished by analyzing polypeptides between the matched samples. The sizes of the proteins in the two tissues are compared, for example, using antibodies of the present invention to detect polypeptides in Western blots of protein extracts from the two tissues. Other changes, such as expression levels and subcellular localization, can also be detected immunologically, using antibodies to the corresponding protein. A higher or lower level of expression of a given polypeptide in a tissue suspected of being diseased, compared with the same protein expression level in a normal tissue, is indicative that the expressed protein has a role in the disease, and provides guidance for preparing a therapeutic agent.

Similarly, comparison of polynucleotide sequences or of gene expression products, *e.g.*, mRNA and protein, between a human tissue that is suspected of being diseased and a normal tissue of a human, are used to follow disease progression or remission in the human. Such comparisons are made as described above. For example, increased or decreased expression of a gene corresponding to an inventive polynucleotide in the tissue suspected of being neoplastic can indicate the presence of neoplastic cells in the tissue. The degree of increased expression of a given gene in the neoplastic tissue relative to expression of the same gene in normal tissue, or differences in the amount of increased expression of a given gene in the neoplastic tissue over time, is used to assess the progression of the neoplasia in that tissue or to monitor the response of the neoplastic tissue to a therapeutic protocol over time.

The expression pattern of any two cell types can be compared, such as low and high metastatic tumor cell lines, malignant or non-malignant cells, or cells from tissue which have and have not been exposed to a therapeutic agent. A genetic predisposition to disease in a human is detected by comparing expression levels of an mRNA or protein corresponding to a polynucleotide of the invention in a fetal tissue with levels associated in normal fetal tissue. Fetal tissues that are used for this purpose include, but are not limited

IOCID: <WO__9938972A2_I_>

to, amniotic fluid, chorionic villi, blood, and the blastomere of an in vitro-fertilized embryo. The comparable normal polynucleotide-related gene is obtained from any tissue. The mRNA or protein is obtained from a normal tissue of a human in which the polynucleotide-related gene is expressed. Differences such as alterations in the nucleotide sequence or size of the same product of the fetal polynucleotide-related gene or mRNA, or alterations in the molecular weight, amino acid sequence, or relative abundance of fetal protein, can indicate a germline mutation in the polynucleotide-related gene of the fetus, which indicates a genetic predisposition to disease. Particular diagnostic and prognostic uses of the disclosed polynucleotides are described in more detail below.

E. Diagnostic, Prognostic, and Other Uses Based On Differential Expression

In general, diagnostic methods of the invention for involve detection of a level or amount of a gene product, particularly a differentially expressed gene product, in a test sample obtained from a patient suspected of having or being susceptible to a disease (*e.g.*, breast cancer, lung cancer, colon cancer and/or metastatic forms thereof), and comparing the detected levels to those levels found in normal cells (*e.g.*, cells substantially unaffected by cancer) and/or other control cells (*e.g.*, to differentiate a cancerous cell from a cell affected by dysplasia). Furthermore, the severity of the disease can be assessed by comparing the detected levels of a differentially expressed gene product with those levels detected in samples representing the levels of differentially gene product associated with varying degrees of severity of disease.

The term "differentially expressed gene" is intended to encompass a polynucleotide that can, for example, include an open reading frame encoding a gene product (*e.g.*, a polypeptide), and/or introns of such genes and adjacent 5' and 3' non-coding nucleotide sequences involved in the regulation of expression, up to about 20 kb beyond the coding region, but possibly further in either direction. The gene can be introduced into an appropriate vector for extrachromosomal maintenance or for integration into a host genome. In general, a difference in expression level associated with a decrease in expression level of at least about 25%, usually at least about 50% to 75%, more usually at least about 90% or more is indicative of a differentially expressed gene of interest, *i.e.*, a gene that is underexpressed or down-regulated in the test sample relative to a control sample. Furthermore, a difference in expression level associated with an increase in

expression of at least about 25%, usually at least about 50% to 75%, more usually at least about 90% and can be at least about 1 ½-fold, usually at least about 2-fold to about 10-fold, and can be about 100-fold to about 1,000-fold increase relative to a control sample is indicative of a differentially expressed gene of interest, *i.e.*, an overexpressed or up-regulated gene.

"Differentially expressed polynucleotide" as used herein means a nucleic acid molecule (RNA or DNA) having a sequence that represents a differentially expressed gene, *e.g.*, the differentially expressed polynucleotide comprises a sequence (*e.g.*, an open reading frame encoding a gene product) that uniquely identifies a differentially expressed gene so that detection of the differentially expressed polynucleotide in a sample is correlated with the presence of a differentially expressed gene in a sample. "Differentially expressed polynucleotides" is also meant to encompass fragments of the disclosed polynucleotides, *e.g.*, fragments retaining biological activity, as well as nucleic acids homologous, substantially similar, or substantially identical (*e.g.*, having about 90% sequence identity) to the disclosed polynucleotides.

Methods of the subject invention useful in diagnosis or prognosis typically involve comparison of the abundance of a selected differentially expressed gene product in a sample of interest with that of a control to determine any relative differences in the expression of the gene product, where the difference can be measured qualitatively and/or quantitatively. Quantitation can be accomplished, for example, by comparing the level of expression product detected in the sample with the amounts of product present in a standard curve. A comparison can be made visually; by using a technique such as densitometry, with or without computerized assistance; by preparing a representative library of cDNA clones of mRNA isolated from a test sample, sequencing the clones in the library to determine that number of cDNA clones corresponding to the same gene product, and analyzing the number of clones corresponding to that same gene product relative to the number of clones of the same gene product in a control sample; or by using an array to detect relative levels of hybridization to a selected sequence or set of sequences, and comparing the hybridization pattern to that of a control. The differences in expression are then correlated with the presence or absence of an abnormal expression pattern. A variety of different methods for determining the nucleic acid abundance in a sample are known to

OCID: <WO___9938972A2_I_>

those of skill in the art, where particular methods of interest include those described in: Pietu *et al. Genome Res.* (1996) 6:492; Zhao *et al.*, *Gene* (1995) *156*:207; Soares , *Curr. Opin. Biotechnol.* (1977) *8*: 542; Raval, *J. Pharmacol Toxicol Methods* (1994) *32*:125; Chalifour *et al.*, *Anal. Biochem* (1994) *216*:299; Stolz *et al.*, *Mol. Biotechnol.* (1996) *6*:225; Hong *et al.*, *Biosci. Reports* (1982) *2*:907; and McGraw, *Anal. Biochem.* (1984) *143*:298. Also of interest are the methods disclosed in WO 97/27317, the disclosure of which is herein incorporated by reference.

In general, diagnostic assays of the invention involve detection of a gene product of a the polynucleotide sequence (*e.g.*, mRNA or polypeptide) that corresponds to a sequence of "SEQ ID NOS:1-5252." The patient from whom the sample is obtained can be apparently healthy, susceptible to disease (*e.g.*, as determined by family history or exposure to certain environmental factors), or can already be identified as having a condition in which altered expression of a gene product of the invention is implicated.

In the assays of the invention, the diagnosis can be determined based on detected gene product expression levels of a gene product encoded by at least one, preferably at least two or more, at least 3 or more, or at least 4 or more of the polynucleotides having a sequence set forth in "SEQ ID NOS:1-5252," and can involve detection of expression of genes corresponding to all of "SEQ ID NOS:1-5252" and/or additional sequences that can serve as additional diagnostic markers and/or reference sequences. Where the diagnostic method is designed to detect the presence or susceptibility of a patient to cancer, the assay preferably involves detection of a gene product encoded by a gene corresponding to a polynucleotide that is differentially expressed in cancer. For example, a higher level of expression of a polynucleotide corresponding to SEQ ID NO:2024 relative to a level associated with a normal sample can indicate the presence of cancer in the patient from whom the sample is derived. In another example, detection of a lower level of a polynucleotide corresponding to SEQ ID NO:590 relative to a normal level is indicative of the presence of cancer in the patient. Further examples of such differentially expressed polynucleotides are described in the Examples below. Given the provided polynucleotides and information regarding their relative expression levels provided herein, assays using such polynucleotides and detection of their expression levels in diagnosis and prognosis will be readily apparent to the ordinarily skilled artisan.

Any of a variety of detectable labels can be used in connection with the various embodiments of the diagnostic methods of the invention. Suitable detectable labels include fluorochromes,(*e.g.* fluorescein isothiocyanate (FITC), rhodamine, Texas Red, phycoerythrin, allophycocyanin, 6-carboxyfluorescein (6-FAM), 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein, 6-carboxy-X-rhodamine (ROX), 6-carboxy-2',4',7',4,7-hexachlorofluorescein (HEX), 5-carboxyfluorescein (5-FAM) or N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA)), radioactive labels, (*e.g.* ^{32}P, ^{35}S, ^{3}H, *etc.*), and the like. The detectable label can involve a two stage systems (*e.g.*, biotin-avidin, hapten-anti-hapten antibody, *etc.*)

Reagents specific for the polynucleotides and polypeptides of the invention, such as antibodies and nucleotide probes, can be supplied in a kit for detecting the presence of an expression product in a biological sample. The kit can also contain buffers or labeling components, as well as instructions for using the reagents to detect and quantify expression products in the biological sample. Exemplary embodiments of the diagnostic methods of the invention are described below in more detail.

Polypeptide detection in diagnosis. In one embodiment, the test sample is assayed for the level of a differentially expressed polypeptide. Diagnosis can be accomplished using any of a number of methods to determine the absence or presence or altered amounts of the differentially expressed polypeptide in the test sample. For example, detection can utilize staining of cells or histological sections with labeled antibodies, performed in accordance with conventional methods. Cells can be permeabilized to stain cytoplasmic molecules. In general, antibodies that specifically bind a differentially expressed polypeptide of the invention are added to a sample, and incubated for a period of time sufficient to allow binding to the epitope, usually at least about 10 minutes. The antibody can be detectably labeled for direct detection (*e.g.*, using radioisotopes, enzymes, fluorescers, chemiluminescers, and the like), or can be used in conjunction with a second stage antibody or reagent to detect binding (*e.g.*, biotin with horseradish peroxidase-conjugated avidin, a secondary antibody conjugated to a fluorescent compound, *e.g.* fluorescein, rhodamine, Texas red, *etc.*). The absence or presence of antibody binding can be determined by various methods, including flow cytometry of dissociated cells, microscopy, radiography, scintillation counting, *etc.* Any suitable alternative methods can

XCID: <WO__9938972A2_I_>

of qualitative or quantitative detection of levels or amounts of differentially expressed polypeptide can be used, for example ELISA, western blot, immunoprecipitation, radioimmunoassay, etc.

In general, the detected level of differentially expressed polypeptide in the test sample is compared to a level of the differentially expressed gene product in a reference or control sample, *e.g.*, in a normal cell (negative control) or in a cell having a known disease state (positive control).

mRNA detection. The diagnostic methods of the invention can also or alternatively involve detection of mRNA encoded by a gene corresponding to a differentially expressed polynucleotides of the invention. Any suitable qualitative or quantitative methods known in the art for detecting specific mRNAs can be used. mRNA can be detected by, for example, *in situ* hybridization in tissue sections, by reverse transcriptase-PCR, or in Northern blots containing poly A+ mRNA. One of skill in the art can readily use these methods to determine differences in the size or amount of mRNA transcripts between two samples. For example, the level of mRNA of the invention in a tissue sample suspected of being cancerous or dysplastic is compared with the expression of the mRNA in a reference sample, *e.g.*, a positive or negative control sample (*e.g.*, normal tissue, cancerous tissue, *etc.*).

Any suitable method for detecting and comparing mRNA expression levels in a sample can be used in connection with the diagnostic methods of the invention (see, *e.g.*, U.S. 5,804,382). For example, mRNA expression levels in a sample can be determined by generation of a library of expressed sequence tags (ESTs) from the sample, where the EST library is representative of sequences present in the sample (Adams, et al., (1991) *Science* 252:1651). Enumeration of the relative representation of ESTs within the library can be used to approximate the relative representation of the gene transcript within the starting sample. The results of EST analysis of a test sample can then be compared to EST analysis of a reference sample to determine the relative expression levels of a selected polynucleotide, particularly a polynucleotide corresponding to one or more of the differentially expressed genes described herein.

Alternatively, gene expression in a test sample can be performed using serial analysis of gene expression (SAGE) methodology (Velculescu et al., *Science* (1995)

270:484). In short, SAGE involves the isolation of short unique sequence tags from a specific location within each transcript. The sequence tags are concatenated, cloned, and sequenced. The frequency of particular transcripts within the starting sample is reflected by the number of times the associated sequence tag is encountered with the sequence population.

Gene expression in a test sample can also be analyzed using differential display (DD) methodology. In DD, fragments defined by specific sequence delimiters (*e.g.*, restriction enzyme sites) are used as unique identifiers of genes, coupled with information about fragment length or fragment location within the expressed gene. The relative representation of an expressed gene with a sample can then be estimated based on the relative representation of the fragment associated with that gene within the pool of all possible fragments. Methods and compositions for carrying out DD are well known in the art, see, *e.g.*, U.S. 5,776,683; and U.S. 5,807,680.

Alternatively, gene expression in a sample using hybridization analysis, which is based on the specificity of nucleotide interactions. Oligonucleotides or cDNA can be used to selectively identify or capture DNA or RNA of specific sequence composition, and the amount of RNA or cDNA hybridized to a known capture sequence determined qualitatively or quantitatively, to provide information about the relative representation of a particular message within the pool of cellular messages in a sample. Hybridization analysis can be designed to allow for concurrent screening of the relative expression of hundreds to thousands of genes by using, for example, array-based technologies having high density formats, including filters, microscope slides, or microchips, or solution-based technologies that use spectroscopic analysis (*e.g.*, mass spectrometry). One exemplary use of arrays in the diagnostic methods of the invention is described below in more detail.

OCID: <WO 9938972A2_I_>

Use of a single gene in diagnostic applications. The diagnostic methods of the invention can focus on the expression of a single differentially expressed gene. For example, the diagnostic method can involve detecting a differentially expressed gene, or a polymorphism of such a gene (*e.g.*, a polymorphism in an coding region or control region), that is associated with disease. Disease-associated polymorphisms can include deletion or truncation of the gene, mutations that alter expression level and/or affect activity of the encoded protein, *etc.*

Changes in the promoter or enhancer sequence that affect expression levels of an differentially gene can be compared to expression levels of the normal allele by various methods known in the art. Methods for determining promoter or enhancer strength include quantitation of the expressed natural protein; insertion of the variant control element into a vector with a reporter gene such as β-galactosidase, luciferase, chloramphenicol acetyltransferase, *etc.* that provides for convenient quantitation; and the like.

A number of methods are available for analyzing nucleic acids for the presence of a specific sequence, *e.g.* a disease associated polymorphism. Where large amounts of DNA are available, genomic DNA is used directly. Alternatively, the region of interest is cloned into a suitable vector and grown in sufficient quantity for analysis. Cells that express a differentially expressed gene can be used as a source of mRNA, which can be assayed directly or reverse transcribed into cDNA for analysis. The nucleic acid can be amplified by conventional techniques, such as the polymerase chain reaction (PCR), to provide sufficient amounts for analysis, and a detectable label can be included in the amplification reaction (*e.g.*, using a detectably labeled primer or detectably labeled oligonucleotides) to facilitate detection. The use of the polymerase chain reaction is described in Saiki, *et al.*, *Science* (1985) *239*:487, and a review of techniques can be found in Sambrook, *et al.*, *Molecular Cloning: A Laboratory Manual*, (1989) pp. 14.2. Alternatively, various methods are known in the art that utilize oligonucleotide ligation as a means of detecting polymorphisms, for examples see Riley *et al.*, *Nucl. Acids Res.* (1990) *18*:2887; and Delahunty *et al.*, *Am. J. Hum. Genet.* (1996) *58*:1239.

The sample nucleic acid, *e.g.* amplified or cloned fragment, is analyzed by one of a number of methods known in the art. The nucleic acid can be sequenced by dideoxy or other methods, and the sequence of bases compared to a selected sequence, *e.g.*, to a wild-

type sequence. Hybridization with the polymorphic or variant sequence can also be used to determine its presence in a sample (*e.g.*, by Southern blot, dot blot, *etc.)*. The hybridization pattern of a polymorphic or variant sequence and a control sequence to an array of oligonucleotide probes immobilized on a solid support, as described in US 5,445,934, or in WO 95/35505, can also be used as a means of identifying polymorphic or variant sequences associated with disease. Single strand conformational polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), and heteroduplex analysis in gel matrices are used to detect conformational changes created by DNA sequence variation as alterations in electrophoretic mobility. Alternatively, where a polymorphism creates or destroys a recognition site for a restriction endonuclease, the sample is digested with that endonuclease, and the products size fractionated to determine whether the fragment was digested. Fractionation is performed by gel or capillary electrophoresis, particularly acrylamide or agarose gels.

Screening for mutations in an differentially expressed gene can be based on the functional or antigenic characteristics of the protein. Protein truncation assays are useful in detecting deletions that can affect the biological activity of the protein. Various immunoassays designed to detect polymorphisms in proteins can be used in screening. Where many diverse genetic mutations lead to a particular disease phenotype, functional protein assays have proven to be effective screening tools. The activity of the encoded protein can be determined by comparison with the wild-type protein.

Pattern matching in diagnosis using arrays. In another embodiment, the diagnostic and/or prognostic methods of the invention involve detection of expression of a selected set of genes in a test sample to produce a test expression pattern (TEP). The TEP is compared to a reference expression pattern (REP), which is generated by detection of expression of the selected set of genes in a reference sample (*e.g.*, a positive or negative control sample). The selected set of genes includes at least one of the genes of the invention, which genes correspond to the polynucleotide sequences of "SEQ ID NOS:1-5252." Of particular interest is a selected set of genes that includes gene differentially expressed in the disease for which the test sample is to be screened.

"Reference sequences" or "reference polynucleotides" as used herein in the context of differential gene expression analysis and diagnosis/prognosis refers to a selected set of

IOCID: <WO__9938972A2_I_>

polynucleotides, which selected set includes at least one or more of the differentially expressed polynucleotides described herein. A plurality of reference sequences, preferably comprising positive and negative control sequences, can be included as reference sequences. Additional suitable reference sequences are found in Genbank, Unigene, and other nucleotide sequence databases (including, *e.g.*, expressed sequence tag (EST), partial, and full-length sequences).

"Reference array" means an array having reference sequences for use in hybridization with a sample, where the reference sequences include all, at least one of, or any subset of the differentially expressed polynucleotides described herein. Usually such an array will include at least 3 different reference sequences, and can include any one or all of the provided differentially expressed sequences. Arrays of interest can further comprise sequences, including polymorphisms, of other genetic sequences, particularly other sequences of interest for screening for a disease or disorder (*e.g.*, cancer, dysplasia, or other related or unrelated diseases, disorders, or conditions). The oligonucleotide sequence on the array will usually be at least about 12 nt in length, and can be of about the length of the provided sequences, or can extend into the flanking regions to generate fragments of 100 nt to 200 nt in length or more.

A "reference expression pattern" or "REP" as used herein refers to the relative levels of expression of a selected set of genes, particularly of differentially expressed genes, that is associated with a selected cell type, *e.g.*, a normal cell, a cancerous cell, a cell exposed to an environmental stimulus, and the like. A "test expression pattern" or "TEP" refers to relative levels of expression of a selected set of genes, particularly of differentially expressed genes, in a test sample (*e.g.*, a cell of unknown or suspected disease state, from which mRNA is isolated).

"Diagnosis" as used herein generally includes determination of a subject's susceptibility to a disease or disorder, determination as to whether a subject is presently affected by a disease or disorder, as well as to the prognosis of a subject affected by a disease or disorder (*e.g.*, identification of pre-metastatic or metastatic cancerous states, stages of cancer, or responsiveness of cancer to therapy). The present invention particularly encompasses diagnosis of subjects in the context of breast cancer (*e.g.*, carcinoma in situ (*e.g.*, ductal carcinoma in situ), estrogen receptor (ER)-positive breast

cancer, ER-negative breast cancer, or other forms and/or stages of breast cancer), lung cancer (*e.g.*, small cell carcinoma, non-small cell carcinoma, mesothelioma, and other forms and/or stages of lung cancer), and colon cancer (*e.g.*, adenomatous polyp, colorectal carcinoma, and other forms and/or stages of colon cancer).

"Sample" or "biological sample" as used throughout here are generally meant to refer to samples of biological fluids or tissues, particularly samples obtained from tissues, especially from cells of the type associated with the disease for which the diagnostic application is designed (*e.g.*, ductal adenocarcinoma), and the like. "Samples" is also meant to encompass derivatives and fractions of such samples (*e.g.*, cell lysates). Where the sample is solid tissue, the cells of the tissue can be dissociated or tissue sections can be analyzed.

REPs can be generated in a variety of ways according to methods well known in the art. For example, REPs can be generated by hybridizing a control sample to an array having a selected set of polynucleotides (particularly a selected set of differentially expressed polynucleotides), acquiring the hybridization data from the array, and storing the data in a format that allows for ready comparison of the REP with a TEP. Alternatively, all expressed sequences in a control sample can be isolated and sequenced, *e.g.*, by isolating mRNA from a control sample, converting the mRNA into cDNA, and sequencing the cDNA. The resulting sequence information roughly or precisely reflects the identity and relative number of expressed sequences in the sample. The sequence information can then be stored in a format (*e.g.*, a computer-readable format) that allows for ready comparison of the REP with a TEP. The REP can be normalized prior to or after data storage, and/or can be processed to selectively remove sequences of expressed genes that are of less interest or that might complicate analysis (*e.g.*, some or all of the sequences associated with housekeeping genes can be eliminated from REP data).

TEPs can be generated in a manner similar to REPs, *e.g.*, by hybridizing a test sample to an array having a selected set of polynucleotides, particularly a selected set of differentially expressed polynucleotides, acquiring the hybridization data from the array, and storing the data in a format that allows for ready comparison of the TEP with a REP. The REP and TEP to be used in a comparison can be generated simultaneously, or the TEP can be compared to previously generated and stored REPs.

DCID: <WO___9938972A2_I_>

In one embodiment of the invention, comparison of a TEP with a REP involves hybridizing a test sample with a reference array, where the reference array has one or more reference sequences for use in hybridization with a sample. The reference sequences include all, at least one of, or any subset of the differentially expressed polynucleotides described herein. Hybridization data for the test sample is acquired, the data normalized, and the produced TEP compared with a REP generated using an array having the same or similar selected set of differentially expressed polynucleotides. Probes that correspond to sequences differentially expressed between the two samples will show decreased or increased hybridization efficiency for one of the samples relative to the other.

Reference arrays can be produced according to any suitable methods known in the art. For example, methods of producing large arrays of oligonucleotides are described in U.S. 5,134,854, and U.S. 5,445,934 using light-directed synthesis techniques. Using a computer controlled system, a heterogeneous array of monomers is converted, through simultaneous coupling at a number of reaction sites, into a heterogeneous array of polymers. Alternatively, microarrays are generated by deposition of pre-synthesized oligonucleotides onto a solid substrate, for example as described in PCT published application no. WO 95/35505.

Methods for collection of data from hybridization of samples with a reference arrays are also well known in the art. For example, the polynucleotides of the reference and test samples can be generated using a detectable fluorescent label, and hybridization of the polynucleotides in the samples detected by scanning the microarrays for the presence of the detectable label. Methods and devices for detecting fluorescently marked targets on devices are known in the art. Generally, such detection devices include a microscope and light source for directing light at a substrate. A photon counter detects fluorescence from the substrate, while an x-y translation stage varies the location of the substrate. A confocal detection device that can be used in the subject methods is described in U.S. Patent no. 5,631,734. A scanning laser microscope is described in Shalon et al., *Genome Res.* (1996) *6*:639. A scan, using the appropriate excitation line, is performed for each fluorophore used. The digital images generated from the scan are then combined for subsequent analysis. For any particular array element, the ratio of the fluorescent signal from one

BNS

sample (*e.g.*, a test sample) is compared to the fluorescent signal from another sample (*e.g.*, a reference sample), and the relative signal intensity determined.

Methods for analyzing the data collected from hybridization to arrays are well known in the art. For example, where detection of hybridization involves a fluorescent label, data analysis can include the steps of determining fluorescent intensity as a function of substrate position from the data collected, removing outliers, *i.e.* data deviating from a predetermined statistical distribution, and calculating the relative binding affinity of the targets from the remaining data. The resulting data can be displayed as an image with the intensity in each region varying according to the binding affinity between targets and probes.

In general, the test sample is classified as having a gene expression profile corresponding to that associated with a disease or non-disease state by comparing the TEP generated from the test sample to one or more REPs generated from reference samples (*e.g.*, from samples associated with cancer or specific stages of cancer, dysplasia, samples affected by a disease other than cancer, normal samples, *etc.*). The criteria for a match or a substantial match between a TEP and a REP include expression of the same or substantially the same set of reference genes, as well as expression of these reference genes at substantially the same levels (*e.g.*, no significant difference between the samples for a signal associated with a selected reference sequence after normalization of the samples, or at least no greater than about 25% to about 40% difference in signal strength for a given reference sequence. In general, a pattern match between a TEP and a REP includes a match in expression, preferably a match in qualitative or quantitative expression level, of at least one of, all or any subset of the differentially expressed genes of the invention.

Pattern matching can be performed manually, or can be performed using a computer program. Methods for preparation of substrate matrices (*e.g.*, arrays), design of oligonucleotides for use with such matrices, labeling of probes, hybridization conditions, scanning of hybridized matrices, and analysis of patterns generated, including comparison analysis, are described in, for example, U.S. 5,800,992.

F. Use of the Polynucleotides of the Invention in Cancer

Oncogenesis involves the unbridled growth, dedifferentiation and abnormal migration of cells. Cancerous cells can have the ability to compress, invade, and destroy

normal tissue. Cancerous cells may also metastasize to other parts of the body via the bloodstream or the lymph system and colonize in these other areas. Different cancers are classified by the cell from which the cancerous cell is derived and from its cellular morphology and/or state of differentiation.

Somatic genetic abnormalities cause cancer initiation and progression. Cancer generally is clonally formed, *i.e.* gain of function of oncogenes and loss of function of tumor suppressor genes within a single cell transform the cell to be cancerous, and that single cell grows and divides to form a cancerous lesion. The genes known to be involved in cancer initiation and progression are involved in numerous cellular functions, including developmental differentiation, cell cycle regulation, cell signaling, immunological response, DNA replication, and DNA repair.

The identification and characterization of genetic or biochemical markers in blood or tissues that will detect the earliest changes along the carcinogenesis pathway and monitor the efficacy of various therapies and preventive interventions is a major goal of cancer research. Scientists have identified genetic changes in stool specimens that indicate the stages of colon cancer, and other biomarkers such as gene mutations, hormone receptors, proteins that inhibit metastasis, and enzymes that metabolize drugs are all being used to determine the severity and predict the course of breast, prostate, lung, and other cancers.

Recent advances in the pathogenesis of certain cancers has been helpful in determining patient treatment. The level of expression of certain polynucleotides can be indicative of a poorer prognosis, and therefore warrant more aggressive chemo- or radio-therapy for a patient. The correlation of novel surrogate tumor specific features with response to treatment and outcome in patients has defined certain prognostic indicators that allow the design of tailored therapy based on the molecular profile of the tumor. These therapies include antibody targeting and gene therapy. Moreover, a promising level of one or more marker polynucleotides can provide impetus for not aggressively treating a particular patient, thus sparing the patient the deleterious side effects of aggressive therapy. Determining expression of certain polynucleotides and comparison of a patients profile with known expression in normal tissue and variants of the disease allows

a determination of the best possible treatment for a patient, both in terms of specificity of treatment and in terms of comfort level of the patient.

Surrogate tumor markers, such as polynucleotide expression, can also be used to better classify, and thus diagnose and treat, different forms and disease states of cancer. Two classifications widely used in oncology that can benefit from identification of the expression levels of the polynucleotides of the invention are staging of the cancerous disorder, and grading the nature of the cancerous tissue.

Staging. Staging is a process used by physicians to describe how advanced the cancerous state is in a patient. Staging assists the physician in determining a prognosis, planning treatment and evaluating the results of such treatment. Different staging systems are used for different types of cancer, but each generally involves the following determinations: the type of tumor, indicated by T; whether the cancer has metastasized to nearby lymph nodes, indicated by N; and whether the cancer has metastasized to more distant parts of the body, indicated by M. This system of staging is called the TNM system. Generally, if a cancer is only detectable in the area of the primary lesion without having spread to any lymph nodes it is called Stage I. If it has spread only to the closest lymph nodes, it is called Stage II. In Stage III, the cancer has generally spread to the lymph nodes in near proximity to the site of the primary lesion. Cancers that have spread to a distant part of the body, such as the liver, bone, brain or another site, are called Stage IV, the most advanced stage.

Currently, the determination of staging is done using pathological techniques and is based more on the presence or absence of malignant tissue rather than the characteristics of the tumor type. Presence or absence of malignant tissue is based primarily on the gross morphology of the cells in the areas biopsied. The polynucleotides of the invention can facilitate fine-tuning of the staging process by identifying markers for the aggresivity of a cancer, *e.g.* the metastatic potential, as well as the presence in different areas of the body. Thus, a Stage II cancer with a polynucleotide signifying a high metastatic potential cancer can be used to change a borderline Stage II tumor to a Stage III tumor, justifying more aggressive therapy. Conversely, the presence of a polynucleotide signifying a lower metastatic potential allows more conservative staging of a tumor.

DOCID: <WO__9938972A2_I_>

Grading of cancers. Grade is a term used to describe how closely a tumor resembles normal tissue of its same type. Based on the microscopic appearance of a tumor, pathologists will identify the grade of a tumor based on parameters such as cell morphology, cellular organization, and other markers of differentiation. As a general rule, the grade of a tumor corresponds to its rate of growth or aggressiveness. That is, undifferentiated or high-grade tumors grow more quickly than well differentiated or low-grade tumors. Information about tumor grade is useful in planning treatment and predicting prognosis.

The American Joint Commission on Cancer has recommended the following guidelines for grading tumors: 1) GX Grade cannot be assessed; 2) G1 Well differentiated; G2 Moderately well differentiated; 3) G3 Poorly differentiated; 4) G4 Undifferentiated. Although grading is used by pathologists to describe most cancers, it plays a more important role in treatment planning for certain types than for others. An example is the Gleason system that is specific for prostate cancer, which uses grade numbers to describe the degree of differentiation. Lower Gleason scores indicate well-differentiated cells. Intermediate scores denote tumors with moderately differentiated cells. Higher scores describe poorly differentiated cells. Grade is also important in some types of brain tumors and soft tissue sarcomas.

The polynucleotides of the invention can be especially valuable in determining the grade of the tumor, as they not only can aid in determining the differentiation status of the cells of a tumor, they can also identify factors other than differentiation that are valuable in determining the aggressivity of a tumor, such as metastatic potential.

Familial Cancer Genes. A number of cancer syndromes are linked to Mendelian inheritance of a predisposition to develop particular cancers. The following table contains a list of cancer types that can be inherited, and for which the gene or genes responsible have been identified. Most of the cancer types listed can occur as part of several different genetic conditions, each caused by alterations in a different gene.

Cancer Type	Genetic Condition	Gene
Brain	Li-Fraumeni syndrome	TP53
Brain	Neurofibromatosis 1	NF1
	Neurofibromatosis 2	NF2
	von Hippel-Lindau syndrome	VHL

Cancer Type	Genetic Condition	Gene
	Tuberous sclerosis 2	TSC2
Breast	Hereditary breast/ovarian cancer 1	BRCA1
	Hereditary breast/ovarian cancer 2	BRCA2
	Li-Fraumeni syndrome	TP53
	Ataxia telangiectasia	ATM
Colon	Familial adenomatous polyposis (FAP)	APC
	Hereditary non-polyposis colon cancer (HNPCC) 1	HMSH2
	Hereditary non-polyposis colon cancer (HNPCC) 2	hMLH1
	Hereditary non-polyposis colon cancer (HNPCC) 3	hPMS1
	Hereditary non-polyposis colon cancer (HNPCC) 4	hPMS2
Endocrine (parathyroid, pituitary, GI endocrine)	Multiple endocrine neoplasia 1 (MEN1)	MEN1
Endocrine (pheochromacytoma, medullary thyroid)	Multiple endocrine neoplasia 2 (MEN2)	RET
Endometrial	Hereditary non-polyposis colon cancer (HNPCC) 1	hMSH2
	Hereditary non-polyposis colon cancer (HNPCC) 2	hMLH1
	Hereditary non-polyposis colon cancer (HNPCC) 3	hPMS1
	Hereditary non-polyposis colon cancer (HNPCC) 4	hPMS2
Eye	Hereditary retinoblastoma	RB1
Hematologic (lymphomas and leukemia)	Li-Fraumeni syndrome	TP53
	Ataxia telangiectasia	ATM
Kidney	Hereditary Wilms' tumor	WT1
	von Hippel-Lindau syndrome	VHL
	Tuberous sclerosis 2	TSC2
Ovary	Hereditary breast/ovarian cancer 1	BRCA1
	Hereditary breast/ovarian cancer 2	BRCA2
Sarcoma	Hereditary retinoblastoma	RB1
	Li-Fraumeni syndrome	TP53
	Neurofibromatosis 1	NF1
Skin	Hereditary melanoma 1	CDKN2
	Hereditary melanoma 2	CDK4
	Basal cell naevus (Gorlin) syndrome	PTCH
Stomach	Hereditary non-polyposis colon cancer (HNPCC) 1	hMSH2
	Hereditary non-polyposis colon cancer (HNPCC) 2	hMLH1
	Hereditary non-polyposis colon cancer (HNPCC) 3	hPMS1
	Hereditary non-polyposis colon cancer (HNPCC) 4	hPMS2

The polynucleotides of the invention can be especially useful to monitor patients having any of the above syndromes to detect potentially malignant events at a molecular level before they are detectable at a gross morphological level. As can be seen from the table, a number of genes are involved in multiple forms of cancer. Thus, a polynucleotide of the invention identified as important for metastatic colon cancer can also have clinical implications for a patient diagnosed with stomach cancer or endometrial cancer.

OCID: <WO___9938972A2_I_>

Lung Cancer. Lung cancer is one of the most common cancers in the United States, accounting for about 15 percent of all cancer cases, or 170,000 new cases each year. At this time, over half of the lung cancer cases in the United States are in men, but the number found in women is increasing and will soon equal that in men. Today more women die of lung cancer than of breast cancer. Lung cancer is especially difficult to diagnose and treat because of the large size of the lungs, which allows cancer to develop for years undetected. In fact, lung cancer can spread outside the lungs without causing any symptoms. Adding to the confusion, the most common symptom of lung cancer, a persistent cough, can often be mistaken for a cold or bronchitis.

Although there are more than a dozen different kinds of lung cancer, the two main types of lung cancer are small cell and nonsmall cell, which encompass about 90% of all lung cancer cases. Small cell carcinoma (also called oat cell carcinoma), which usually starts in one of the larger bronchial tubes, grows fairly rapidly, and is likely to be large by the time of diagnosis. Nonsmall cell lung cancer (NSCLC) is made up of three general subtypes of lung cancer. Epidermoid carcinoma (also called squamous cell carcinoma) usually starts in one of the larger bronchial tubes and grows relatively slowly. The size of these tumors can range from very small to quite large. Adenocarcinoma starts growing near the outside surface of the lung and can vary in both size and growth rate. Some slowly growing adenocarcinomas are described as alveolar cell cancer. Large cell carcinoma starts near the surface of the lung, grows rapidly, and the growth is usually fairly large when diagnosed. Other less common forms of lung cancer are carcinoid, cylindroma, mucoepidermoid, and malignant mesothelioma.

Currently, CT scans, MRIs, X-rays, sputum cytology, and biopsies are used to diagnose nonsmall cell lung cancer. The form and cellular origin of the lung cancer is diagnosed primarily through biopsy from either a surgical biopsy or a needle aspiration of lung tissue, and usually the biopsy is prompted from an abnormality identified on an X-ray. In some cases, sputum cytology can reveal lung cancers in patients with normal X-rays or can determine the type of lung cancer, but because it cannot pinpoint the tumor's location, a positive sputum cytology test is usually followed by further tests. Since these tests are based in large part on gross morphology of the tissue, the diagnosis of a particular kind of tumor is largely subjective, and the diagnosis can vary significantly between clinicians.

The polynucleotides of the invention can be used to distinguish types of lung cancer as well as identifying traits specific to a certain patient's cancer. For example, if the patient's biopsy expresses a polynucleotide that is associated with a low metastatic potential, it may justify leaving a larger portion of the patient's lung in surgery to remove the lesion. Alternatively, a smaller lesion with expression of a polynucleotide that is associated with high metastatic potential may justify a more radical removal of lung tissue and/or the surrounding lymph nodes, even if no metastasis can be identified through pathological examination.

Similarly, the expression of polynucleotides of the invention can be used in the diagnosis, prognosis and management of colorectal cancer. The differential expression of a polynucleotide in hyperplasia can be used as a diagnostic marker for metastatic lung cancer. The polynucleotides of the invention that would be especially useful for this purpose are those that exhibit differential expression between high metastatic versus low metastatic lung cancer , *i.e.* SEQ ID NOS: 174, 254, 466, 571, 574, 590, 922, 1355, 1422, 2007, 2038, 2245, 10, 54, 65, 171, 203, 252, 253, 285, 419, 420, 491, 525, 526, 552, 693, 700, 726, 742, 746, 861, 990, 1088, 1288, 1417, 1444, 1454, 1570, 1597, 1979, 2024, 2034, and 2126. Detection of malignant lung cancer with a higher metastatic potential can be determined using expression levels of any of these sequences alone or in combination with the levels of expression of other known genes.

Breast Cancer. The National Cancer Institute (NCI) estimates that about 1 in 8 women in the United States will develop breast cancer during her lifetime. Clinical breast examination and mammography are recommended as combined modalities for breast cancer screening, and the nature of the cancer will often depend upon the location of the tumor and the cell type from which the tumor is derived. The majority of breast cancers are adenocarcinomas subtypes, which can be summarized as follows:

Ductal carcinoma in situ (DCIS): Ductal carcinoma in situ is the most common type of noninvasive breast cancer. In DCIS, the malignant cells have not metastasized through the walls of the ducts into the fatty tissue of the breast. Comedocarcinoma is a type of DCIS that is more likely than other types of DCIS to come back in the same area after lumpectomy. It is more closely linked to eventual development of invasive ductal carcinoma than other forms of DCIS.

OCID: <WO___9938972A2_I_>

Infiltrating (or invasive) ductal carcinoma (IDC): this type of cancer has metastasized through the wall of the duct and invaded the fatty tissue of the breast. At this point, it has the potential to use the lymphatic system and bloodstream for metastasis to more distant parts of the body. Infiltrating ductal carcinoma accounts for about 80% of breast cancers.

Lobular carcinoma in situ (LCIS): While not a true cancer, LCIS (also called lobular neoplasia) is sometimes classified as a type of noninvasive breast cancer. It does not penetrate through the wall of the lobules. Although it does not itself usually become an invasive cancer, women with this condition have a higher risk of developing an invasive breast cancer in the same breast, or in the opposite breast.

Infiltrating (or invasive) lobular carcinoma (ILC): ILC is similar to IDC, in that it has the potential metastasize elsewhere in the body. About 10% to 15% of invasive breast cancers are invasive lobular carcinomas. ILC can be more difficult to detect by mammogram than IDC.

Inflammatory breast cancer: This rare type of invasive breast cancer accounts for about 1% of all breast cancers and is extremely aggressive. Multiple skin symptoms associated with this cancer are caused by cancer cells blocking lymph vessels or channels in the skin over the breast.

Medullary carcinoma: This special type of infiltrating breast cancer has a relatively well defined, distinct boundary between tumor tissue and normal tissue. It accounts for about 5% of breast cancers. The prognosis for this kind of breast cancer is better than for other types of invasive breast cancer.

Mucinous carcinoma: This rare type of invasive breast cancer originates from mucus-producing cells. The prognosis for mucinous carcinoma is better than for the more common types of invasive breast cancer.

Paget's disease of the nipple: This type of breast cancer starts in the ducts and spreads to the skin of the nipple and the areola. It is a rare type of breast cancer, occurring in only 1% of all cases. Paget's disease can be associated with in situ carcinoma, or with infiltrating breast carcinoma. If no lump can be felt in the breast tissue, and the biopsy shows DCIS but no invasive cancer, the prognosis is excellent.

Phyllodes tumor: This very rare type of breast tumor forms from the stroma of the breast, in contrast to carcinomas which develop in the ducts or lobules. Phyllodes (also spelled phylloides) tumors are usually benign, but are malignant on rare occasions. Nevertheless, malignant phyllodes tumors are very rare and less than 10 women per year in the US die of this disease. Benign phyllodes tumors are successfully treated by removing the mass and a narrow margin of normal breast tissue.

Tubular carcinoma: Accounting for about 2% of all breast cancers, tubular carcinomas are a special type of infiltrating breast carcinoma. They have a better prognosis than usual infiltrating ductal or lobularcarcinomas.

High-quality mammography combined with clinical breast exam remains the only screening method clearly tied to reduction in breast cancer mortality. Lower dose x-rays, digitized computer rather than film images, and the use of computer programs to assist diagnosis, are almost ready for widespread dissemination. Other technologies also are being developed, including magnetic resonance imaging and ultrasound. In addition, a very low radiation exposure technique, positron emission tomography has the potential for detecting early breast cancer.

It is also possible to differentiate between non-cancerous breast tissue and malignant breast tissue by analyzing differential gene expression between tissues. In addition, there may be several possible alterations that lead to the various possible types of breast cancer. The different types of breast tumors (*e.g.*, invasive vs. non-invasive, ductal vs. axillary lymph node) can be differentiable from one another by the identification of the differences in genes expressed by different types of breast tumor tissues (Porter-Jordan *et al.*, *Hematol Oncol Clin North Am* (1994) *8*:73). Breast cancer can thus be generally diagnosed by detection of expression of a gene or genes associated with breast tumors. Where enough information is available about the differential gene expression between various types of breast tumor tissues, the specific type of breast tumor can also be diagnosed.

For example, increased estrogen receptor (ER) expression in normal breast epithileum, while not itself indicative of malignant tissue, is a known risk marker for development of breast cancer. Khan SA *et al.*, *Cancer Res* (1994) *54*:993. Malignant breast cancer is often divided into two groups, ER-positive and ER-negative, based on the

IOCID: <WO___9938972A2_I_>

estrogen receptor status of the tissue. The ER status represents different survival length and response to hormone therapy, and is thought to represent either: 1) an indicator of different stages of the disease, or 2) an indicator that allows differentiation between two similar but distinct diseases. K. Zhu *et al.*, *Med. Hypoth.* (1997) *49*:69. A number of other genes are known to vary expression between either different stages of cancer or different types of similar breast cancer.

Similarly, the expression of polynucleotides of the invention can be used in the diagnosis and management of breast cancer. The differential expression of a polynucleotide in human breast tumor tissue can be used as a diagnostic marker for human breast cancer. The polynucleotides of the invention that would be especially useful for this purpose are those that exhibit differential expression between breast cancer tissue with a high metastatic potential and a low metastatic potential, *i.e.* SEQ ID NOS:15, 36, 44, 89, 172, 203, 261, 419, 420, 503, 552, 564, 570, 590, 693, 707, 711, 726, 746, 756, 990, 1122, 1142, 1286, 1289, 1435, 1860, 1933, 1934, 1979, 1980, 2007, 2023, 2409, 2486, 45, 146, 154, 159, 165, 174, 183, 364, 366, 387, 496, 510, 512, 529, 560, 606, 644, 646, 754, 875, 902, 921, 942, 1095, 1104, 1131, 1170, 1184, 1205, 1354, 1387, 1535, 1751, 1764, 1777, 1795, 1869, 1882, 1890, 1915, 2040, 2059, 2223, 2245, 2300, 2325, 2462, 2488, 2492; Detection of breast cancer can be determined using expression levels of any of these sequences alone or in combination. Determination of the aggressive nature and/or the metastatic potential of a breast cancer can also be determined by comparing levels of one or more polynucleotides of the invention and comparing levels of another sequence known to vary in cancerous tissue, *e.g.* ER expression. In addition, development of breast cancer can be detected by examining the ratio of SEQ ID NO: to the levels of steroid hormones (*e.g.*, testosterone or estrogen) or to other hormones (*e.g.*, growth hormone, insulin). Thus expression of specific marker polynucleotides can be used to discriminate between normal and cancerous breast tissue, to discriminate between breast cancers with different cells of origin, to discriminate between breast cancers with different potential metastatic rates, etc.

Diagnosis of breast cancer can also involve comparing the expression of a polynucleotide of the invention with the expression of other sequences in non-malignant breast tissue samples in comparison to one or more forms of the diseased tissue. A comparison of expression of one or more polynucleotides of the invention between the

samples provides information on relative levels of these polynucleotides as well as the ratio of these polynucleotides to the expression of other sequences in the tissue of interest compared to normal.

This risk of breast cancer is elevated significantly by the presence of an inherited risk for breast cancer, such as a mutation in BRCA-1 or BRCA-2. New diagnostic tools are being developed to address the needs of higher risk patients to complement mammography and physical examinations for early detection of breast cancer, particularly among younger women. The presence of antigen or expression markers in nipple aspirate fluid (NAF) samples collected from one or both breasts can be useful for useful for risk assessment or early cancer detection. Breast cytology and biomarkers obtained by random fine needle aspiration have been used to identify hyperplasia with atypia and overexpression of p53 and EGFR. The polynucleotides of the invention can be used in multivariate analysis with expression studies with genes such as p53 and EGFR as risk predictors and as surrogate endpoint biomarkers for breast cancer.

As well as being used for diagnosis and risk assessment, the expression of certain genes can also correlated to prognosis of a disease state. The expression of particular gene have been used as prognostic indicators for breast cancer including increased expression of *c-erbB-2*, pS2, ER, progesterone receptor, epidermal growth factor receptor (EGFR), *neu*, *myc*, *bcl-2*, *int2*, cytosolic tyrosine kinase, cyclin E, *prad-1*, *hst*, uPA, PAI-1, PAI-2, cathepsin D, as well as the presence of a number of cancer-specific antigens, *e.g.* CEA, CA M26, CA M29 and CA 15.3. Davis, *Br. J. Biomed Sci.* (1996) *53*:157. Poor prognosis has also been linked to a decrease in expression of certain genes, such as *p53, Rb, nm23.* The expression of the polynucleotides of the invention can be of prognostic value for determining the metastatic potential of a malignant breast cancer, as this molecules are differentially expressed between high and low metastatic potential tissues tumors. The levels of these polynucleotides in patients with malignant breast cancer can compared to normal tissue, malignant tissue with a known high potential metastatic level, and malignant tissue with a known lower level of metastatic potential to provide a prognosis for a particular patient. Such a prognosis is predictive of the extent and nature of the cancer. The determined prognosis is useful in determining the prognosis of a patient with breast cancer, both for initial treatment of the disease and for longer-term monitoring of the same

OCID: <WO___9938972A2_I_>

patient. If samples are taken from the same individual over a period of time, differences in polynucleotide expression that are specific to that patient can be identified and closely watched.

Colon Cancer. Colorectal cancer is one of the most common neoplasms in humans and perhaps the most frequent form of hereditary neoplasia. Prevention and early detection are key factors in controlling and curing colorectal cancer. Indeed, colorectal cancer is the second most preventable cancer, after lung cancer. Colorectal cancer begins as polyps, which are small, benign growths of cells that form on the inner lining of the colon. Over a period of several years, some of these polyps accumulate additional mutations and become cancerous. About 20 percent of all cases of colon cancer are thought to be related to heredity. Currently, multiple familial colorectal cancer disorders have been identified, which are summarized as follows:

Familial adenomatous polyposis (FAP): This condition results in a person having hundreds or even thousands of polyps in the colon and rectum that usually first appear during the teenage years. Cancer nearly always develops in one or more of these polyps between the ages of 30 and 50.

Gardner's syndrome: Like FAP, Gardner's syndrome results in polyps and colorectal cancers that develop at a young age. It can also cause benign tumors of the skin, soft connective tissue and bones.

Hereditary nonpolyposis colon cancer (HNPCC): People with this condition tend to develop colorectal cancer at a young age, without first having many polyps. HNPCC has an autosomal dominant pattern of inheritance with variable but high penetrance estimated to be about 90%. HNPCC underlies 0.5%-10% of all cases of colorectal cancer. An understanding of the mechanisms behind the development of HNPCC is emerging, and genetic presymptomatic testing, now being conducted in research settings, soon will be available on a widespread basis for individuals identified at risk for this disease.

Familial colorectal cancer in Ashkenazi Jews: Recent research has found an inherited tendency to developing colorectal cancer among some Jews of Eastern European descent. Like people with FAP, Gardner's syndrome, and HNPCC, their increased risk is due to an inherited mutation present in about 6% of American Jews.

Several tests are currently used to screen for colorectal cancer, including digital rectal examination, fecal occult blood test, sigmoidoscopy, colonoscopy, virtual colonoscopy and MRI. Each of these tests identifies potential colorectal cancer lesions, or a risk of development of these lesions, at a fairly gross morphological level.

The sequential alteration of a number of genes is associated with malignant adenocarcinoma, including the genes DCC, p53, ras, and FAP. For a review, see *e.g.* Fearon ER, *et al.*, *Cell* (1990) *61(5)*:759; Hamilton SR *et al.*, *Cancer* (1993) *72*:957; Bodmer W, *et al.*, *Nat Genet.* (1994) *4(3)*:217; Fearon ER, *Ann N Y Acad Sci.* (1995) *768*:101. Molecular genetic alterations are thus promising as potential diagnostic and prognostic indicators in colorectal carcinoma and molecular genetics of colorectal carcinoma since it is possible to differentiate between different types of colorectal neoplasias using molecular markers. Colorectal cancer can thus be generally diagnosed by detection of expression of a gene or genes associated with colorectal tumors.

Similarly, the expression of polynucleotides of the invention can be used in the diagnosis, prognosis and management of colorectal cancer. The differential expression of a polynucleotide in hyperplasia can be used as a diagnostic marker for colon cancer. The polynucleotides of the invention that would be especially useful for this purpose are those that exhibit differential expression between malignant metastatic colon cancer and normal patient tissue , *i.e.* SEQ ID NOS:228, 280, 355, 491, 603, 680, 752, 753, 1241, 1264, 1401, 1442, 1514, 1851, 1915, 2024, 2066, 33, 250, 282, 370, 387, 443, 460, 545, 560, 703, 704, 1095, 1104, 1205, 1354, 1387, 1734, 1742, 1954, 2262, 2325, 1899, 252, 253, 491, 581, 693, 726, 746, 1780, 1899, 65, 252, 253, 581, 693, 716, 726, 746, 1780, 1899, and 1780. Detection of malignant colon cancer can be determined using expression levels of any of these sequences alone or in combination with the levels of expression.

Determination of the aggressive nature and/or the metastatic potential of a colon cancer can also be determined by comparing levels of one or more polynucleotides of the invention and comparing total levels of another sequence known to vary in cancerous tissue, *e.g.* p53 expression. In addition, development of colon cancer can be detected by examining the ratio of any of the polynucleotides of the invention to the levels of oncogenes (*e.g.* ras) or tumor suppressor genes (*e.g.* FAP or p53). Thus expression of specific marker polynucleotides can be used to discriminate between normal and cancerous

XCID: <WO___9938972A2_I_>

breast tissue, to discriminate between breast cancers with different cells of origin, to discriminate between breast cancers with different potential metastatic rates, etc.

G. Use of Polynucleotides to Screen for Peptide Analogs and Antagonists

Polypeptides encoded by the instant polynucleotides and corresponding full length genes can be used to screen peptide libraries to identify binding partners, such as receptors, from among the encoded polypeptides.

A library of peptides can be synthesized following the methods disclosed in U.S. Pat. No. 5,010,175 ('175), and in WO 91/17823. As described below in brief, one prepares a mixture of peptides, which is then screened to identify the peptides exhibiting the desired signal transduction and receptor binding activity. In the '175 method, a suitable peptide synthesis support (*e.g.*, a resin) is coupled to a mixture of appropriately protected, activated amino acids. The concentration of each amino acid in the reaction mixture is balanced or adjusted in inverse proportion to its coupling reaction rate so that the product is an equimolar mixture of amino acids coupled to the starting resin. The bound amino acids are then deprotected, and reacted with another balanced amino acid mixture to form an equimolar mixture of all possible dipeptides. This process is repeated until a mixture of peptides of the desired length (*e.g.*, hexamers) is formed. Note that one need not include all amino acids in each step: one can include only one or two amino acids in some steps (*e.g.*, where it is known that a particular amino acid is essential in a given position), thus reducing the complexity of the mixture. After the synthesis of the peptide library is completed, the mixture of peptides is screened for binding to the selected polypeptide. The peptides are then tested for their ability to inhibit or enhance activity. Peptides exhibiting the desired activity are then isolated and sequenced.

The method described in WO 91/17823 is similar. However, instead of reacting the synthesis resin with a mixture of activated amino acids, the resin is divided into twenty equal portions (or into a number of portions corresponding to the number of different amino acids to be added in that step), and each amino acid is coupled individually to its portion of resin. The resin portions are then combined, mixed, and again divided into a number of equal portions for reaction with the second amino acid. In this manner, each reaction can be easily driven to completion. Additionally, one can maintain separate "subpools" by treating portions in parallel, rather than combining all resins at each step.

This simplifies the process of determining which peptides are responsible for any observed receptor binding or signal transduction activity.

In such cases, the subpools containing, *e.g.*, 1-2,000 candidates each are exposed to one or more polypeptides of the invention. Each subpool that produces a positive result is then resynthesized as a group of smaller subpools (sub-subpools) containing, *e.g.*, 20-100 candidates, and reassayed. Positive sub-subpools can be resynthesized as individual compounds, and assayed finally to determine the peptides that exhibit a high binding constant. These peptides can be tested for their ability to inhibit or enhance the native activity. The methods described in WO 91/7823 and U.S. Patent No. 5,194,392 (herein incorporated by reference) enable the preparation of such pools and subpools by automated techniques in parallel, such that all synthesis and resynthesis can be performed in a matter of days.

Peptide agonists or antagonists are screened using any available method, such as signal transduction, antibody binding, receptor binding, mitogenic assays, chemotaxis assays, etc. The methods described herein are presently preferred. The assay conditions ideally should resemble the conditions under which the native activity is exhibited *in vivo*, that is, under physiologic pH, temperature, and ionic strength. Suitable agonists or antagonists will exhibit strong inhibition or enhancement of the native activity at concentrations that do not cause toxic side effects in the subject. Agonists or antagonists that compete for binding to the native polypeptide can require concentrations equal to or greater than the native concentration, while inhibitors capable of binding irreversibly to the polypeptide can be added in concentrations on the order of the native concentration.

The end results of such screening and experimentation will be at least one novel polypeptide binding partner, such as a receptor, encoded by a gene or a cDNA corresponding to a polynucleotide of the invention, and at least one peptide agonist or antagonist of the novel binding partner. Such agonists and antagonists can be used to modulate, enhance, or inhibit receptor function in cells to which the receptor is native, or in cells that possess the receptor as a result of genetic engineering. Further, if the novel receptor shares biologically important characteristics with a known receptor, information about agonist/antagonist binding can facilitate development of improved agonists/antagonists of the known receptor.

XOCID: <WO__9938972A2_I_>

H. Pharmaceutical Compositions and Therapeutic Uses

Pharmaceutical compositions can comprise polypeptides, antibodies, or polynucleotides of the claimed invention. The pharmaceutical compositions will comprise a therapeutically effective amount of either polypeptides, antibodies, or polynucleotides of the claimed invention.

The term "therapeutically effective amount" as used herein refers to an amount of a therapeutic agent to treat, ameliorate, or prevent a desired disease or condition, or to exhibit a detectable therapeutic or preventative effect. The effect can be detected by, for example, chemical markers or antigen levels. Therapeutic effects also include reduction in physical symptoms, such as decreased body temperature. The precise effective amount for a subject will depend upon the subject's size and health, the nature and extent of the condition, and the therapeutics or combination of therapeutics selected for administration. Thus, it is not useful to specify an exact effective amount in advance. However, the effective amount for a given situation is determined by routine experimentation and is within the judgment of the clinician. For purposes of the present invention, an effective dose will generally be from about 0.01 mg/ kg to 50 mg/kg or 0.05 mg/kg to about 10 mg/kg of the DNA constructs in the individual to which it is administered.

A pharmaceutical composition can also contain a pharmaceutically acceptable carrier. The term "pharmaceutically acceptable carrier" refers to a carrier for administration of a therapeutic agent, such as antibodies or a polypeptide, genes, and other therapeutic agents. The term refers to any pharmaceutical carrier that does not itself induce the production of antibodies harmful to the individual receiving the composition, and which can be administered without undue toxicity. Suitable carriers can be large, slowly metabolized macromolecules such as proteins, polysaccharides, polylactic acids, polyglycolic acids, polymeric amino acids, amino acid copolymers, and inactive virus particles. Such carriers are well known to those of ordinary skill in the art.

Pharmaceutically acceptable salts can be used therein, for example, mineral acid salts such as hydrochlorides, hydrobromides, phosphates, sulfates, and the like; and the salts of organic acids such as acetates, propionates, malonates, benzoates, and the like. A thorough discussion of pharmaceutically acceptable excipients is available in *Remington's Pharmaceutical Sciences* (Mack Pub. Co., N.J. 1991).

Pharmaceutically acceptable carriers in therapeutic compositions can include liquids such as water, saline, glycerol and ethanol. Auxiliary substances, such as wetting or emulsifying agents, pH buffering substances, and the like, can also be present in such vehicles. Typically, the therapeutic compositions are prepared as injectables, either as liquid solutions or suspensions; solid forms suitable for solution in, or suspension in, liquid vehicles prior to injection can also be prepared. Liposomes are included within the definition of a pharmaceutically acceptable carrier.

Delivery Methods. Once formulated, the compositions of the invention can be (1) administered directly to the subject (*e.g.*, as polynucleotide or polypeptides); (2) delivered ex vivo, to cells derived from the subject (*e.g.*, as in *ex vivo* gene therapy); or (3) delivered *in vitro* for expression of recombinant proteins (*e.g.*, polynucleotides). Direct delivery of the compositions will generally be accomplished by injection, either subcutaneously, intraperitoneally, intravenously or intramuscularly, or delivered to the interstitial space of a tissue. The compositions can also be administered into a tumor or lesion. Other modes of administration include oral and pulmonary administration, suppositories, and transdermal applications, needles, and gene guns or hyposprays. Dosage treatment can be a single dose schedule or a multiple dose schedule.

Methods for the ex vivo delivery and reimplantation of transformed cells into a subject are known in the art and described in *e.g.*, International Publication No. WO 93/14778. Examples of cells useful in ex vivo applications include, for example, stem cells, particularly hematopoetic, lymph cells, macrophages, dendritic cells, or tumor cells. Generally, delivery of nucleic acids for both ex vivo and in vitro applications can be accomplished by, for example, dextran-mediated transfection, calcium phosphate precipitation, polybrene mediated transfection, protoplast fusion, electroporation, encapsulation of the polynucleotide(s) in liposomes, and direct microinjection of the DNA into nuclei, all well known in the art.

Once a gene corresponding to a polynucleotide of the invention has been found to correlate with a proliferative disorder, such as neoplasia, dysplasia, and hyperplasia, the disorder can be amenable to treatment by administration of a therapeutic agent based on the provided polynucleotide or corresponding polypeptide.

DOCID: <WO___9938972A2_I_>

Preparation of antisense polynucleotides is discussed above. Neoplasias that are treated with the antisense composition include, but are not limited to, cervical cancers, melanomas, colorectal adenocarcinomas, Wilms' tumor, retinoblastoma, sarcomas, myosarcomas, lung carcinomas, leukemias, such as chronic myelogenous leukemia, promyelocytic leukemia, monocytic leukemia, and myeloid leukemia, and lymphomas, such as histiocytic lymphoma. Proliferative disorders that are treated with the therapeutic composition include disorders such as anhydric hereditary ectodermal dysplasia, congenital alveolar dysplasia, epithelial dysplasia of the cervix, fibrous dysplasia of bone, and mammary dysplasia. Hyperplasias, for example, endometrial, adrenal, breast, prostate, or thyroid hyperplasias or pseudoepitheliomatous hyperplasia of the skin, are treated with antisense therapeutic compositions based upon a polynucleotide of the invention. Even in disorders in which mutations in the corresponding gene are not implicated, downregulation or inhibition of expression of a gene corresponding to a polynucleotide of the invention can have therapeutic application. For example, decreasing gene expression can help to suppress tumors in which enhanced expression of the gene is implicated.

Both the dose of the antisense composition and the means of administration are determined based on the specific qualities of the therapeutic composition, the condition, age, and weight of the patient, the progression of the disease, and other relevant factors. Administration of the therapeutic antisense agents of the invention includes local or systemic administration, including injection, oral administration, particle gun or catheterized administration, and topical administration. Preferably, the therapeutic antisense composition contains an expression construct comprising a promoter and a polynucleotide segment of at least 12, 22, 25, 30, or 35 contiguous nucleotides of the antisense strand of a polynucleotide disclosed herein. Within the expression construct, the polynucleotide segment is located downstream from the promoter, and transcription of the polynucleotide segment initiates at the promoter.

Various methods are used to administer the therapeutic composition directly to a specific site in the body. For example, a small metastatic lesion is located and the therapeutic composition injected several times in several different locations within the body of tumor. Alternatively, arteries which serve a tumor are identified, and the therapeutic composition injected into such an artery, in order to deliver the composition directly into

the tumor. A tumor that has a necrotic center is aspirated and the composition injected directly into the now empty center of the tumor. The antisense composition is directly administered to the surface of the tumor, for example, by topical application of the composition. X-ray imaging is used to assist in certain of the above delivery methods.

Receptor-mediated targeted delivery of therapeutic compositions containing an antisense polynucleotide, subgenomic polynucleotides, or antibodies to specific tissues is also used. Receptor-mediated DNA delivery techniques are described in, for example, Findeis *et al., Trends Biotechnol.* (1993) *11*:202; Chiou *et al., Gene Therapeutics: Methods And Applications Of Direct Gene Transfer* (J.A. Wolff, ed.) (1994); Wu *et al., J. Biol. Chem.* (1988) *263*:621; Wu *et al., J. Biol. Chem.* (1994) *269*:542; Zenke *et al., Proc. Natl. Acad. Sci. (USA)* (1990) *87*:3655; Wu *et al., J. Biol. Chem.* (1991) *266*:338. Preferably, receptor-mediated targeted delivery of therapeutic compositions containing antibodies of the invention is used to deliver the antibodies to specific tissue.

Therapeutic compositions containing antisense subgenomic polynucleotides are administered in a range of about 100 ng to about 200 mg of DNA for local administration in a gene therapy protocol. Concentration ranges of about 500 ng to about 50 mg, about 1 μg to about 2 mg, about 5 μg to about 500 μg, and about 20 μg to about 100 μg of DNA can also be used during a gene therapy protocol. Factors such as method of action and efficacy of transformation and expression are considerations which will affect the dosage required for ultimate efficacy of the antisense subgenomic polynucleotides. Where greater expression is desired over a larger area of tissue, larger amounts of antisense subgenomic polynucleotides or the same amounts readministered in a successive protocol of administrations, or several administrations to different adjacent or close tissue portions of, for example, a tumor site, may be required to effect a positive therapeutic outcome. In all cases, routine experimentation in clinical trials will determine specific ranges for optimal therapeutic effect. A more complete description of gene therapy vectors, especially retroviral vectors, is contained in U.S. Serial No. 08/869,309, which is expressly incorporated herein, and in section G below.

For polynucleotide-related genes encoding polypeptides or proteins with anti-inflammatory activity, suitable use, doses, and administration are described in U.S. Patent No. 5,654,173. Therapeutic agents also include antibodies to proteins and polypeptides

XOCID: <WO__9938972A2_I_>

encoded by the polynucleotides of the invention and related genes, as described in U.S. Patent No. 5,654,173.

I. Gene Therapy

The therapeutic polynucleotides and polypeptides of the present invention can be utilized in gene delivery vehicles. The gene delivery vehicle can be of viral or non-viral origin (see generally, Jolly, *Cancer Gene Therapy* (1994) *1*:51; Kimura, *Human Gene Therapy* (1994) *5*:845; Connelly, *Human Gene Therapy* (1995) *1*:185; and Kaplitt, *Nature Genetics* (1994) *6*:148). Gene therapy vehicles for delivery of constructs including a coding sequence of a therapeutic of the invention can be administered either locally or systemically. These constructs can utilize viral or non-viral vector approaches. Expression of such coding sequences can be induced using endogenous mammalian or heterologous promoters. Expression of the coding sequence can be either constitutive or regulated.

The present invention can employ recombinant retroviruses which are constructed to carry or express a selected nucleic acid molecule of interest. Retrovirus vectors that can be employed include those described in EP 0 415 731; WO 90/07936; WO 94/03622; WO 93/25698; WO 93/25234; U.S. Patent No. 5, 219,740; WO 93/11230; WO 93/10218; Vile and Hart, *Cancer Res.* (1993) *53*:3860; Vile *et al.*, *Cancer Res.* (1993) *53*:962; Ram et al., *Cancer Res.* (1993) *53*:83; Takamiya *et al.*, J. *Neurosci. Res.* (1992) *33*:493; Baba et al., *J. Neurosurg.* (1993) *79*:729; U.S. Patent No. 4,777,127; GB Patent No. 2,200,651; and EP 0 345 242. Preferred recombinant retroviruses include those described in WO 91/02805.

Packaging cell lines suitable for use with the above-described retroviral vector constructs can be readily prepared (see, *e.g.*, WO 95/30763 and WO 92/05266), and used to create producer cell lines (also termed vector cell lines) for the production of recombinant vector particles. Within particularly preferred embodiments of the invention, packaging cell lines are made from human (such as HT1080 cells) or mink parent cell lines, thereby allowing production of recombinant retroviruses that can survive inactivation in human serum.

The present invention also employs alphavirus-based vectors that can function as gene delivery vehicles. Such vectors can be constructed from a wide variety of alphaviruses, including, for example, Sindbis virus vectors, Semliki forest virus (ATCC VR-67; ATCC VR-1247), Ross River virus (ATCC VR-373; ATCC VR-1246) and

Venezuelan equine encephalitis virus (ATCC VR-923; ATCC VR-1250; ATCC VR 1249; ATCC VR-532). Representative examples of such vector systems include those described in U.S. Patent Nos. 5,091,309; 5,217,879; and 5,185,440; WO 92/10578; WO 94/21792; WO 95/27069; WO 95/27044; and WO 95/07994. Gene delivery vehicles of the present invention can also employ parvovirus such as adeno-associated virus (AAV) vectors. Representative examples include the AAV vectors disclosed by Srivastava in WO 93/09239, Samulski et al., *J. Virol.* (1989) *63*:3822; Mendelson et al., *Virol.* (1988) *166*:154; and Flotte *et al.*, *PNAS* (1993) *90*:10613.

Representative examples of adenoviral vectors include those described by Berkner, *Biotechniques* (1988) *6*:616; Rosenfeld *et al.*, *Science* (1991) *252*:431; WO 93/19191; Kolls *et al.*, *PNAS* (1994) *91*:215; Kass-Eisler *et al.*, *PNAS* (1993) *90*:11498; Guzman *et al.*, *Circulation* (1993) *88*:2838; Guzman *et al.*., *Cir. Res.* (1993) *73*:1202; Zabner *et al.*, *Cell* (1993) *75*:207; Li *et al.*, *Hum. Gene Ther.* (1993) *4*:403; Cailaud *et al.*, *Eur. J. Neurosci.* (1993) *5*:1287; Vincent *et al.*, *Nat. Genet.* (1993) *5*:130; Jaffe *et al.*, *Nat. Genet.* (1992) *1*:372; and Levrero *et al.*, *Gene* (1991) *101*:195. Exemplary adenoviral gene therapy vectors employable in this invention also include those described in WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655. Administration of DNA linked to killed adenovirus as described in Curiel, *Hum. Gene Ther.* (1992) *3*:147 can be employed.

Other gene delivery vehicles and methods can be employed, including polycationic condensed DNA linked or unlinked to killed adenovirus alone, for example Curiel, *Hum. Gene Ther.* (1992) *3*:147; ligand linked DNA, for example see Wu, *J. Biol. Chem.* (1989) *264*:16985; eukaryotic cell delivery vehicles cells, for example see U.S. Pat. No. 5,814,482; WO 95/07994; WO 96/17072; WO 95/30763; and WO 97/42338; deposition of photopolymerized hydrogel materials; hand-held gene transfer particle gun, as described in U.S. Patent No. 5,149,655; ionizing radiation as described in U.S. Patent No. 5,206,152 and in WO92/11033; nucleic charge neutralization or fusion with cell membranes. Additional approaches are described in Philip, *Mol. Cell Biol.* (1994) *14*:2411, and in Woffendin, *Proc. Natl. Acad. Sci.* (1994) *91*:1581.

Naked DNA can also be employed. Exemplary naked DNA introduction methods are described in WO 90/11092 and U.S. Patent No. 5,580,859. Liposomes that can act as

OCID: <WO___9938972A2_I_>

gene delivery vehicles are described in U.S. Patent No. 5,422,120; WO 95/13796; WO 94/23697; WO 91/14445; and EP 0524968.

Further non-viral delivery suitable for use includes mechanical delivery systems such as the approach described in Woffendin *et al., Proc. Natl. Acad. Sci. USA* (1994) *91*(24):11581. Moreover, the coding sequence and the product of expression of such can be delivered through deposition of photopolymerized hydrogel materials. Other conventional methods for gene delivery that can be used for delivery of the coding sequence include, for example, use of hand-held gene transfer particle gun, as described in U.S. Patent No. 5,149,655; use of ionizing radiation for activating transferred gene, as described in U.S. Patent No. 5,206,152 and WO 92/11033.

The present invention will now be illustrated by reference to the following examples which set forth particularly advantageous embodiments. However, it should be noted that these embodiments are illustrative and are not to be construed as restricting the invention in any way.

EXAMPLES

Example 1: Source of Biological Materials and Overview of Novel Polynucleotides Expressed by the Biological Materials

Human colon cancer cell line Km12L4-A (Morika, W. A. K. et al., *Cancer Research* (1988) *48*:6863) was used to construct a cDNA library from mRNA isolated from the cells. As described in the above overview, a total of 4,693 sequences expressed by the Km12L4-A cell line were isolated and analyzed; most sequences were about 275-300 nucleotides in length. The KM12L4-A cell line is derived from the KM12C cell line. The KM12C cell line, which is poorly metastatic (low metastatic) was established in culture from a Dukes' stage B_2 surgical specimen (Morikawa *et al. Cancer Res.* (1988) *48*:6863). The KML4-A is a highly metastatic subline derived from KM12C (Yeatman *et al. Nucl. Acids. Res.* (1995) *23*:4007; Bao-Ling *et al. Proc. Annu. Meet. Am. Assoc. Cancer. Res.* (1995) *21*:3269). The KM12C and KM12C-derived cell lines (*e.g.*, KM12L4, KM12L4-A, *etc.*) are well-recognized in the art as a model cell line for the study of colon cancer (see, *e.g.*, Moriakawa *et al.*, *supra*; Radinsky *et al. Clin. Cancer Res.* (1995) *1*:19; Yeatman *et al.*, (1995) *supra*; Yeatman *et al. Clin. Exp. Metastasis* (1996) *14*:246).

The sequences were first masked to eliminate low complexity sequences using the XBLAST masking program (Claverie "Effective Large-Scale Sequence Similarity Searches," In: Computer Methods for Macromolecular Sequence Analysis, Doolittle, ed., *Meth. Enzymol.* 266:212-227 Academic Press, NY, NY (1996); see particularly Claverie, in "Automated DNA Sequencing and Analysis Techniques" Adams *et al.*, eds., Chap. 36, p. 267 Academic Press, San Diego, 1994 and Claverie *et al. Comput. Chem.* (1993) 17:191). Generally, masking does not influence the final search results, except to eliminate sequences of relative little interest due to their low complexity, and to eliminate multiple "hits" based on similarity to repetitive regions common to multiple sequences, e.g., Alu repeats. Masking resulted in the elimination of 43 sequences. The remaining sequences were then used in a BLASTN vs. Genbank search with search parameters of greater than 70% overlap, 99% identity, and a p value of less than 1×10^{-40}, which search resulted in the discarding of 1,432 sequences. Sequences from this search also were discarded if the inclusive parameters were met, but the sequence was ribosomal or vector-derived.

The resulting sequences from the previous search were classified into three groups (1, 2 and 3 below) and searched in a BLASTX vs. NRP (non-redundant proteins) database

XOCID: <WO__9938972A2_I_>

search: (1) unknown (no hits in the Genbank search), (2) weak similarity (greater than 45% identity and p value of less than 1 x 10^{-5}), and (3) high similarity (greater than 60% overlap, greater than 80% identity, and p value less than 1 x 10^{-5}). This search resulted in discard of 98 sequences as having greater than 70% overlap, greater than 99% identity, and p value of less than 1 x 10^{-40}.

The remaining sequences were classified as unknown (no hits), weak similarity, and high similarity (parameters as above). Two searches were performed on these sequences. First, a BLAST vs. EST database search resulted in discard of 1771 sequences (sequences with greater than 99% overlap, greater than 99% similarity and a p value of less than 1 x 10^{-40}; sequences with a p value of less than 1 x 10^{-65} when compared to a database sequence of human origin were also excluded). Second, a BLASTN vs. Patent GeneSeq database resulted in discard of 15 sequences (greater than 99% identity; p value less than 1 x 10^{-40}; greater than 99% overlap).

The remaining sequences were subjected to screening using other rules and redundancies in the dataset. Sequences with a p value of less than 1 x 10^{-111} in relation to a database sequence of human origin were specifically excluded. The final result provided the 2502 sequences listed in the accompanying Sequence Listing. The Sequence Listing is arranged beginning with sequences with no similarity to any sequence in a database searched, and ending with sequences with the greatest similarity. Each identified polynucleotide represents sequence from at least a partial mRNA transcript. Polynucleotides that were determined to be novel were assigned a sequence identification number.

The novel polynucleotides were assigned sequence identification numbers SEQ ID NOS:1-2502. The DNA sequences corresponding to the novel polynucleotides are provided in the Sequence Listing. The majority of the sequences are presented in the Sequence Listing in the 5' to 3' direction. A small number of sequences are listed in the Sequence Listing in the 5' to 3' direction but the sequence as written is actually 3' to 5'. These sequences are readily identified with the designation "AR" in the Sequence Name in Table 1 (inserted before the claims). The sequences correctly listed in the 5' to 3' direction in the Sequence Listing are designated "AF." Table 1 provides: 1) the SEQ ID NO assigned to each sequence for use in the present specification; 2) the filing date of the U.S. priority application in which the sequence was first filed; 3) the SEQ ID NO assigned to the sequence in the priority application; 4) the sequence name used as an internal identifier of

the sequence; 5) the name assigned to the clone from which the sequence was isolated; and 6) the number of the cluster to which the sequence is assigned (Cluster ID; where the cluster ID is 0, the sequence was not assigned to any cluster

Because the provided polynucleotides represent partial mRNA transcripts, two or more polynucleotides of the invention may represent different regions of the same mRNA transcript and the same gene. Thus, if two or more SEQ ID NOS: are identified as belonging to the same clone, then either sequence can be used to obtain the full-length mRNA or gene. In addition, some sequences are identified with multiple SEQ ID NOS, since these sequences were present in more than one filing. For example, SEQ ID NO:87 and SEQ ID NO:1000 represent the same sequence.

In order to confirm the sequences of SEQ ID NOS:1-2502, inserts of the clones corresponding to these polynucleotides were re-sequenced. These "validation" sequences are provided in SEQ ID NOS:2503-5106. Of these validation sequences, SEQ ID NOS:3040, 3545, 3863, 4511, 4726, and 4749 are not true validation sequences. Instead, SEQ ID NOS:3545, 4511, 4726, and 4749 represent "placeholder" sequences, *i.e.*, sequences that were inserted into the Sequence Listing only to prevent renumbering of the subsequent sequences during generation of the Sequence Listing. Thus, reference to "SEQ ID NOS:1-5252," "SEQ ID NOS:1-5106," or other ranges of SEQ ID NOS that include these placeholder sequences should be read to exclude SEQ ID NOS:3545, 4511, 4726, and 4749.

The validation sequences were often longer than the original polynucleotide sequences they validate, and thus often provide additional sequence information. Validation sequences can be correlated with the original sequences they validate by referring to Table 1. For example, validation sequences of SEQ ID NOS:2503-3039, 3041-3544, 3546-3862 3864-4510, and 4512-4725 share the clone name of the sequence of SEQ ID NOS:1-2502 that they validate.

Example 2: Results of Public Database Search to Identify Function of Gene Products

SEQ ID NOS:1-2502, as well as the validation sequences SEQ ID NOS:2503-3039, 3041-3544, 3546-3862 3864-4510, and 4512-4725 xx:clf were translated in all three reading frames to determine the best alignment with the individual sequences. These amino acid sequences and nucleotide sequences are referred, generally, as query sequences, which are aligned with the individual sequences. Query and individual sequences were

OCID: <WO 9938972A2 I >

aligned using the BLAST programs, available over the world wide web at http://ww.ncbi.nlm.nih.gov/BLAST/. Again the sequences were masked to various extents to prevent searching of repetitive sequences or poly-A sequences, using the XBLAST program for masking low complexity as described above in Example 1.

Table 2 (inserted before the claims) shows the results of the alignments. Table 2 refers to each sequence by its SEQ ID NO:, the accession numbers and descriptions of nearest neighbors from the Genbank and Non-Redundant Protein searches, and the p values of the search results.

For each of "SEQ ID NOS:1-5106," the best alignment to a protein or DNA sequence is included in Table 2. The activity of the polypeptide encoded by "SEQ ID NOS:1-5106" is the same or similar to the nearest neighbor reported in Table 2. The accession number of the nearest neighbor is reported, providing a reference to the activities exhibited by the nearest neighbor. The search program and database used for the alignment also are indicated as well as a calculation of the p value.

Full length sequences or fragments of the polynucleotide sequences of the nearest neighbors can be used as probes and primers to identify and isolate the full length sequence of "SEQ ID NOS:1-5106." The nearest neighbors can indicate a tissue or cell type to be used to construct a library for the full-length sequences of "SEQ ID NOS:1-5106."

"SEQ ID NOS:1-5106" and the translations thereof may be human homologs of known genes of other species or novel allelic variants of known human genes. In such cases, these new human sequences are suitable as diagnostics or therapeutics. As diagnostics, the human sequences "SEQ ID NOS:1-5106" exhibit greater specificity in detecting and differentiating human cell lines and types than homologs of other species. The human polypeptides encoded by "SEQ ID NOS1-5106" are likely to be less immunogenic when administered to humans than homologs from other species. Further, on administration to humans, the polypeptides encoded by "SEQ ID NOS:1-5106" can show greater specificity or can be better regulated by other human proteins than are homologs from other species.

Example 3: Members of Protein Families

The validation sequences ("SEQ ID NOS:2503-5106") were used to conduct a profile search as described in the specification above. Several of the polynucleotides of the invention were found to encode polypeptides having characteristics of a polypeptide

belonging to a known protein families (and thus represent new members of these protein families) and/or comprising a known functional domain (Table 3, inserted prior to claims). Thus the invention encompasses fragments, fusions, and variants of such polynucleotides that retain biological activity associated with the protein family and/or functional domain identified herein.

Start and stop indicate the position within the individual sequences that align with the query sequence having the indicated SEQ ID NO. The direction (Dir) indicates the orientation of the query sequence with respect to the individual sequence, where forward (for) indicates that the alignment is in the same direction (left to right) as the sequence provided in the Sequence Listing and reverse (rev) indicates that the alignment is with a sequence complementary to the sequence provided in the Sequence Listing.

Some polynucleotides exhibited multiple profile hits because, for example, the particular sequence contains overlapping profile regions, and/or the sequence contains two different functional domains. These profile hits are described in more detail below. The acronyms used in Table 3 are provided in parentheses following the full name of the protein family or functional domain to which they refer.

a) Seven Transmembrane Integral Membrane Proteins -- Rhodopsin Family (7tm_1). Several of the validation sequences, and thus their corresponding sequence within SEQ ID NOS:1-2502, correspond to a sequence encoding a polypeptide that is a member of the seven transmembrane receptor rhodopsin family. G-protein coupled receptors of the seven transmembrane rhodopsin family (also called R7G) are an extensive group of hormones, neurotransmitters, and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins (Strosberg A.D. *Eur. J. Biochem.* (1991) *196*:1, Kerlavage A.R. *Curr. Opin. Struct. Biol.* (1991) *1*:394, Probst, et al., *DNA Cell Biol.* (1992) *11*:1, Savarese, et al., Biochem. J. (1992) *283*:1, http://www.gcrdb.uthscsa.edu/, http://swift.embl-heidelberg.de/7tm/. The receptors that are currently known to belong to this family are: 1) 5-hydroxytryptamine (serotonin) 1A to 1F, 2A to 2C, 4, 5A, 5B, 6 and 7 (Branchek T., *Curr. Biol.* (1993) *3*:315); 2) acetylcholine, muscarinic-type, M1 to M5; 3) adenosine A1, A2A, A2B and A3 (Stiles G.L. *J. Biol. Chem.* (1992) *267*:6451; 4) adrenergic alpha-1A to -1C; alpha-2A to -2D; beta-1 to -3 (Friell T. et al., *Trends Neurosci.* (1988) *11*:321); 5) angiotensin II types I and II; 6) bombesin subtypes 3 and 4; 7) bradykinin B1 and B2; 8) c3a and C5a anaphylatoxin; 9) cannabinoid CB1 and CB2; 10) chemokines C-C CC-CKR-1 to CC-CKR-8; 11)

XCID: <WO___9938972A2_I_>

Chemokines C-X-C CXC-CKR-1 to CXC-CKR-4; 12) Cholecystokinin-A and cholecystokinin-B/gastrin Dopamine D1 to D5 (Stevens C.F., *Curr. Biol.* (1991) *1*:20); 13) Endothelin ET-a and ET-b (Sakurai T. et al., *Trends Pharmacol. Sci.* (1992) *13*:103-107); 14) fMet-Leu-Phe (fMLP) (Nformyl peptide); 15) Follicle stimulating hormone (FSH-R); 16) Galanin; 17) Gastrin-releasing peptide (GRP-R); 18) Gonadotropin-releasing hormone (GNRH-R); 19) Histamine H1 and H2 (gastric receptor I); 20) Lutropin-choriogonadotropic hormone (LSH-R) (Salesse R., et al., *Biochimie* (1991) *73*:109); 21) Melanocortin MC1R to MC5R; 22) Melatonin; 23) Neuromedin B (NMB-R); 24) Neuromedin K (NK-3R); 25) Neuropeptide Y types 1 to 6; 26) Neurotensin (NT-R); 27) Octopamine (tyramine), from insects; 28) Odorants (Lancet D., et al., *Curr. Biol.* (1993)*3*:668; 29) Opioids delta-, kappa- and mu-types (Uhl G.R., et al., *Trends Neurosci.* (1994) *17*:89; 30) Oxytocin (OT-R); 31) Platelet activating factor (PAF-R); 32) Prostacyclin; 33) Prostaglandin D2; 34) Prostaglandin E2, EP1 to EP4 subtypes; 35) Prostaglandin F2; 36) Purinoreceptors (ATP) (Barnard E.A., et al., *Trends Pharmacol. Sci.* (1994)*15*:67; 37); Somatostatin types 1 to 5; 38) Substance-K (NK-2R); Substance-P (NK-1R); 39) Thrombin; 40) Thromboxane A2; 41) Thyrotropin (TSH-R) (Salesse R., et al., *Biochimie* (1991) *73*:109); 42) Thyrotropin releasing factor (TRH-R); 42) Vasopressin V1a, V1b and V2; 43) Visual pigments (opsins and rhodopsin) (Applebury M.L., et al., *Vision Res.* (1986) *26*:1881; 44) Proto-oncogene mas; 45) A number of orphan receptors (whose ligand is not known) from mammals and birds; 46) Caenorhabditis elegans putative receptors C06G4.5, C38C10.1, C43C3.2; 47) T27D1.3 and ZC84.4; 48) Three putative receptors encoded in the genome of cytomegalovirus: US27, US28, and UL33; and 49) ECRF3, a putative receptor encoded in the genome of herpesvirus saimiri.

The structure of these receptors is thought to be identical. They have seven hydrophobic regions, each of which most probably spans the membrane. The N-terminus is located on the extracellular side of the membrane and is often glycosylated, while the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three intracellular loops to link the seven transmembrane regions. Most, but not all of these receptors, lack a signal peptide. The most conserved parts of these proteins are the transmembrane regions and the first two cytoplasmic loops. A conserved acidic-Arg-aromatic triplet is present in the N-terminal extremity of the second cytoplasmic loop (Attwood T.K., Eliopoulos E.E., Findlay J.B.C. *Gene* (1991) *98*:153-159) and could be implicated in the interaction with G proteins.

A consensus pattern that contains the conserved triplet and that also spans the major part of the third transmembrane helix is used to detect this widespread family of proteins: [GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM].

b) Seven Transmembrane Integral Membrane Proteins -- Secretin Family (7tm_2). Several of the validation sequences, and thus their corresponding sequence within SEQ ID NOS:1-2502, correspond to a sequence encoding a polypeptide that is a member of the seven transmembrane receptor secretin family. A number of peptide hormones bind to G-protein coupled receptors that, while structurally similar to the majority of G-protein coupled receptors (R7G) (see profile for 7 transmembrane receptors (rhodopsin family), do not show any similarity at the level of their sequence, thus new family whose current known members (Jueppner et al. *Science* (1991) *254*:1024; Hamann et al. *Genomics* (1996) *32*:144).are: 1) calcitonin receptor, 2) calcitonin gene-related peptide receptor; 3) corticotropin releasing factor receptor types 1 and 2; 4) gastric inhibitory polypeptide receptor; 5) glucagon receptor; 6) glucagon-like peptide 1 receptor; 7) growth hormone-releasing hormone receptor; 7) parathyroid hormone / parathyroid hormone-related peptide types 1 and 2; 8) pituitary adenylate cyclase activating polypeptide receptor; 9) secretin receptor; 10) vasoactive intestinal peptide receptor types 1 and 2; 10) insects diuretic hormone receptor; 11) Caenorhabditis elegans putative receptor C13B9.4; 12) Caenorhabditis elegans putative receptor ZK643.3; 13) human leucocyte CD97 (which contains 3 EGF-like domains in its N-terminal section); 14) human cell surface glycoprotein EMR1 (which contains 6 EGF-like domains in it N-terminal section); and 15) mouse cell surface glycoprotein F4/80 (which contains 7 EGF-like domains in its N-terminal section). All of 1) through 10) are coupled to G-proteins which activate both adenylyl cyclase and the phosphatidylinositol-calcium pathway.

Like classical R7G the secretin family of 7 transmembrane proteins contain seven transmembrane regions. Their N-terminus is located on the extracellular side of the membrane and potentially glycosylated, while their C-terminus is cytoplasmic. But apart from these topological similarities they do not share any region of sequence similarity and are therefore probably not evolutionary related.

Every receptor in the 7 transmember secretin family is encoded on multiple exons, and several of these functionally distinct products. The N-terminal extracellular domain of these receptors contains five conserved cysteines residues that may be involved in disulfide

JCID: <WO__9938972A2_I_> BNS page 85

bonds, with a consensus pattern in the region that spans the first three cysteines. One of the most highly conserved regions spans the C-terminal part of the last transmembrane region and the beginning of the adjacent intracellular region. This second region is used as a second signature pattern. The two consensus patterns are:

1) C-x(3)-[FYWLIV]-D-x(3,4)-C-[FW]-x(2)-[STAGV]-x(8,9)-C-[PF]

2) Q-G-[LMFCA]-[LIVMFT]-[LIV]-x-[LIVFST]-[LIF]-[VFYH]-C- [LFY]-x-N-x(2)-V

c) Ank Repeats (ANK). SEQ IS NO:2656, and thus its corresponding sequence within SEQ ID NOS:1-2502, represents a polynucleotide encoding an Ank repeat-containing protein. The ankyrin motif is a 33 amino acid sequence named after the protein ankyrin which has 24 tandem 33-amino-acid motifs. Ank repeats were originally identified in the cell-cycle-control protein cdc10 (Breeden *et al., Nature* (1987) *329*:651). Proteins containing ankyrin repeats include ankyrin, myotropin, I-kappaB proteins, cell cycle protein cdc10, the Notch receptor (Matsuno *et al., Development* (1997) *124(21)*:4265); G9a (or BAT8) of the class III region of the major histocompatibility complex (Biochem J. 290:811-818, 1993), FABP, GABP, 53BP2, Lin12, glp-1, SW14, and SW16. The functions of the ankyrin repeats are compatible with a role in protein-protein interactions (Bork, *Proteins* (1993) *17(4)*:363; Lambert and Bennet, *Eur. J. Biochem.* (1993) *211*:1; Kerr *et al., Current Op. Cell Biol.* (1992) *4*:496; Bennet *et al., J. Biol. Chem.* (1980) *255*:6424).

The 90 kD N-terminal domain of ankyrin contains a series of 24 33-amino-acid ank repeats. (Lux *et al., Nature* (1990) *344*:36-42, Lambert *et al., PNAS USA* (1990) *87*:1730.) The 24 ank repeats form four folded subdomains of 6 repeats each. These four repeat subdomains mediate interactions with at least 7 different families of membrane proteins. Ankyrin contains two separate binding sites for anion exchanger dimers. One site utilizes repeat subdomain two (repeats 7-12) and the other requires both repeat subdomains 3 and 4 (repeats 13-24). Since the anion exchangers exist in dimers, ankyrin binds 4 anion exchangers at the same time (Michaely and Bennett, *J. Biol. Chem.* (1995) *270(37)*:22050). The repeat motifs are involved in ankyrin interaction with tubulin, spectrin, and other membrane proteins. (Lux *et al., Nature* (1990) *344*:36.)

The Rel/NF-kappaB/Dorsal family of transcription factors have activity that is controlled by sequestration in the cytoplasm in association with inhibitory proteins referred to as I-kappaB. (Gilmore, *Cell* (1990) *62*:841; Nolan and Baltimore, *Curr Opin Genet Dev.* (1992) *2*:211; Baeuerle, *Biochim Biophys Acta* (1991) *1072*:63; Schmitz *et al., Trends Cell*

Biol. (1991) *1*:130.) I-kappaB proteins contain 5 to 8 copies of 33 amino acid ankyrin repeats and certain NF-kappaB/rel proteins are also regulated by cis-acting ankyrin repeat containing domains including p105NF-kappaB which contains a series of ankyrin repeats (Diehl and Hannink, *J. Virol.* (1993) *67(12)*:7161). The I-kappaBs and Cactus (also containing ankyrin repeats) inhibit activators through differential interactions with the Rel-homology domain. The gene family includes proto-oncogenes, thus broadly implicating I-kappaB in the control of both normal gene expression and the aberrant gene expression that makes cells cancerous. (Nolan and Baltimore, *Curr Opin Genet Dev.* (1992) *2(2)*:211-220). In the case of rel/NF-kappaB and pp40/I-kappaB(, both the ankyrin repeats and the carboxy-terminal domain are required for inhibiting DNA-binding activity and direct association of pp40/I-kappaB(with rel/NF-kappaB protein. The ankyrin repeats and the carboxy-terminal of pp40/I-kappaB(form a structure that associates with the rel homology domain to inhibit DNA binding activity (Inoue *et al.*, *PNAS USA* (1992) *89*:4333).

The 4 ankyrin repeats in the amino terminus of the transcription factor subunit GABP are required for its interaction with the GABP subunit to form a functional high affinity DNA-binding protein. These repeats can be crosslinked to DNA when GABP is bound to its target sequence. (Thompson *et al.*, *Science* (1991) *253*:762; LaMarco *et al.*, *Science* (1991) *253*:789). Myotrophin, a 12.5 kDa protein having a key role in the initiation of cardiac hypertrophy, comprises ankyrin repeats. The ankyrin repeats are characteristic of a hairpin-like protruding tip followed by a helix-turn-helix motif. The V-shaped helix-turn-helix of the repeats stack sequentially in bundles and are stabilized by compact hydrophobic cores, whereas the protruding tips are less ordered.

d) Eukaryotic Aspartyl Proteases (asp). Several of the validation sequences, and thus their corresponding sequence within SEQ ID NOS:1-2502, correspond to a sequence encoding a novel eukaryotic aspartyl protease. Aspartyl proteases, known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes (Foltmann B., *Essays Biochem.* (1981) *17*:52; Davies D.R., *Annu. Rev. Biophys. Chem.* (1990) *19*:189; Rao J.K.M., *et al.*, *Biochemistry* (1991) *30*:4663) known to exist in vertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases include: 1) Vertebrate gastric pepsins A and C (also known as

XCID: <WO___9938972A2_I_>

gastricsin); 2) Vertebrate chymosin (rennin), involved in digestion and used for making cheese; 3) Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34); 4) Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma; 5) Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21); and 6) Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases; 7) Yeast barrierpepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone; and 8) Fission yeast sxa1 which is involved in degrading or processing the mating pheromones.

Most retroviruses and some plant viruses, such as badnaviruses, encode for an aspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of a polyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gag polyprotein. Because the sequence around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases is conserved, a single signature pattern can be used to identify members of both groups of proteases. The consensus pattern is: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]- x-[LIVMFSTNC]-x-[LIVMFGTA], where D is the active site residue.

e) ATPases Associated with Various Cellular Activities (ATPases). Several of the validation sequences, and thus their corresponding sequence within SEQ ID NOS:1-2502, correspond to a sequence that encodes a novel member of the "ATPases Associated with diverse cellular Activities" (AAA) protein family. The AAA protein family is composed of a large number of ATPases that share a conserved region of about 220 amino acids that contains an ATP-binding site (Froehlich *et al., J. Cell Biol.* (1991) *114*:443; Erdmann *et al. Cell* (1991) *64*:499; Peters *et al., EMBO J.* (1990) *9*:1757; Kunau *et al., Biochimie* (1993) *75*:209-224; Confalonieri *et al., BioEssays* (1995) *17*:639; http://yeamob.pci.chemie.uni-tuebingen.de/AAA/Description.html). The proteins that belong to this family either contain one or two AAA domains.

Proteins containing two AAA domains include: 1) Mammalian and drosophila NSF (N-ethylmaleimide-sensitive fusion protein) and the fungal homolog, SEC18, which are

involved in intracellular transport between the endoplasmic reticulum and Golgi, as well as between different Golgi cisternae; 2) Mammalian transitional endoplasmic reticulum ATPase (previously known as p97 or VCP), which is involved in the transfer of membranes from the endoplasmic reticulum to the golgi apparatus. This ATPase forms a ring-shaped homooligomer composed of six subunits. The yeast homolog, CDC48, plays a role in spindle pole proliferation; 3) Yeast protein PAS1 essential for peroxisome assembly and the related protein PAS1 from Pichia pastoris; 4) Yeast protein AFG2; 5) Sulfolobus acidocaldarius protein SAV and Halobacterium salinarium cdcH, which may be part of a transduction pathway connecting light to cell division.

Proteins containing a single AAA domain include: 1) Escherichia coli and other bacteria ftsH (or hflB) protein. FtsH is an ATP-dependent zinc metallopeptidase that degrades the heat-shock sigma-32 factor, and is an integral membrane protein with a large cytoplasmic C-terminal domain that contain both the AAA and the protease domains; 2) Yeast protein YME1, a protein important for maintaining the integrity of the mitochondrial compartment. YME1 is also a zinc-dependent protease; 3) Yeast protein AFG3 (or YTA10). This protein also contains an AAA domain followed by a zinc-dependent protease domain; 4) Subunits from regulatory complex of the 26S proteasome (Hilt *et al., Trends Biochem. Sci.* (1996) *21*:96), which is involved in the ATP-dependent degradation of ubiquitinated proteins, which subunits include: a) Mammalian 4 and homologs in other higher eukaryotes, in yeast (gene YTA5) and fission yeast (gene mts2); b) Mammalian 6 (TBP7) and homologs in other higher eukaryotes and in yeast (gene YTA2); c) Mammalian subunit 7 (MSS1) and homologs in other higher eukaryotes and in yeast (gene CIM5 or YTA3); d) Mammalian subunit 8 (P45) and homologs in other higher eukaryotes and in yeast (SUG1 or CIM3 or TBY1) and fission yeast (gene let1); e) Other probable subunits include human TBP1, which influences HIV gene expression by interacting with the virus tat transactivator protein, and yeast YTA1 and YTA6; 5) Yeast protein BCS1, a mitochondrial protein essential for the expression of the Rieske iron-sulfur protein; 6) Yeast protein MSP1, a protein involved in intramitochondrial sorting of proteins; 7) Yeast protein PAS8, and the corresponding proteins PAS5 from Pichia pastoris and PAY4 from Yarrowia lipolytica; 8) Mouse protein SKD1 and its fission yeast homolog (SpAC2G11.06); 9) Caenorhabditis elegans meiotic spindle formation protein mei-1; 10) Yeast protein SAP1' 11) Yeast protein YTA7; and 12) Mycobacterium leprae hypothetical protein A2126A.

DOCID: <WO___9938972A2_I_>

In general, the AAA domains in these proteins act as ATP-dependent protein clamps(Confalonieri *et al.* (1995) *BioEssays 17*:639). In addition to the ATP-binding 'A' and 'B' motifs, which are located in the N-terminal half of this domain, thereis a highly conserved region located in the central part of the domain which was used in the development of the signature pattern. The consensus pattern is: [LIVMT]-x-[LIVMT]-[LIVMF]-x-[GATMC]-[ST]-[NS]-x(4)-[LIVM]- D-x-A-[LIFA]-x-R.

f) Bcl-2 family (Bcl-2). SEQ ID NO:3404, and thus the corresponding sequence it validates, represents a polynucleotide encoding an apoptosis regulator protein of the Bcl-2 family. Active cell suicide (apoptosis) is induced by events such as growth factor withdrawal and toxins. It is controlled by regulators, which have either an inhibitory effect on programmed cell death (anti-apoptotic) or block the protective effect of inhibitors (pro-apoptotic) (Vaux, 1993, Curr. Biol. 3:877-878, and White, 1996, Genes Dev. 10:2859-2869). Many viruses have found a way of countering defensive apoptosis by encoding their own anti-apoptosis genes, preventing their target cells from dying prematurely.

All proteins belonging to the Bcl-2 family (Reed et al., 1996, Adv. Exp. Med. Biol. 406:99-112) contain either a BH1, BH2, BH3, or BH4 domain. All anti-apoptotic proteins contain BH1 and BH2 domains; some of them contain an additional N-terminal BH4 domain (Bcl-2, Bcl-x(L), Bcl-w), which is never seen in pro-apoptotic proteins, except for Bcl-x(S). On the other hand, all pro-apoptotic proteins contain a BH3 domain (except for Bad) necessary for dimerization with other proteins of Bcl-2 family and crucial for their killing activity; some of them also contain BH1 and BH2 domains (Bax, Bak). The BH3 domain is also present in some anti-apoptotic protein, such as Bcl-2 or Bcl-x(L). Proteins that are known to contain these domains are listed below.

1. Vertebrate protein Bcl-2. Bcl-2 blocks apoptosis; it prolongs the survival of hematopoietic cells in the absence of required growth factors and also in the presence of various stimuli inducing cellular death. Two isoforms of bcl-2 (alpha and beta) are generated by alternative splicing. Bcl-2 is expressed in a wide range of tissues at various times during development. It forms heterodimers with the Bax proteins.

2. Vertebrate protein Bcl-x. Two isoforms of Bcl-x (Bcl-x(L) and Bcl-x(S)) are generated by alternative splicing. While the longer product (Bcl-x(L)) can protect a growth-factor-dependent cell line from apoptosis, the shorter form blocks the protective effect of Bcl-2 and Bcl-x(L) and acts as an anti-anti-apoptosis protein.

3. Mammalian protein Bax. Bax blocks the anti-apoptosis ability of Bcl-2 with which

it forms heterodimers. There is no evidence that Bax has any activity in the absence of Bcl-2. Three isoforms of bax (alpha, beta and gamma) are generated by alternative splicing.

4. Mammalian protein Bak, which promotes cell death and counteracts the protection from apoptosis provided by Bcl-2.

5. Mammalian protein Bcl-w, which promotes cell survival.

6. Mammalian protein bad, which promotes cell death, and counteracts the protection from apoptosis provided by Bcl-x(L), but not that of Bcl-2.

7. Human protein Bik, which promotes cell death, but cannot counteract the protection from apoptosis provided by Bcl-2.

8. Mouse protein Bid, which induces caspases and apoptosis, and counteracts the protection from apoptosis provided by Bcl-2.

9. Human induced myeloid leukemia cell differentiation protein MCL1. MCL1 is probably involved in programming of differentiation and concomitant maintenance of viability but not proliferation. Its expression increases early during phorbol ester induced differentiation in myeloid leukemia cell line ML-1.

10. Mouse hemopoietic-specific early response protein A1.

11. Mammalian activator of apoptosis Harakiri (Inohara et al., 1997, EMBO J. 16:1686-1694) (also known as neuronal death protein Dp5). This is a small protein of 92 residues that activates apoptosis. It contains a BH3 domain, but no BH1, BH2 or BH4 domains.

The following consensus patterns have been developed for the four BH domains:

1) [LVME]-[FT]-x-[GSD]-[GL]-x(1,2)-[NS]-[YW]-G-R-[LIV]- [LIVC]-[GAT]-[LIVMF](2)-x-F-[GSAE]-[GSARY]

2) W-[LIM]-x(3)-[GR]-G-[WQ]-[DENSAV]-x-[FLGA]-[LIVFTC]

3) [LIVAT]-x(3)-L-[KARQ]-x-[IVAL]-G-D-[DESG]-[LIMFV]-[DENSHQ]-[LVSHRQ]-[NSR]

4) [DS]-[NT]-R-[AE]-[LI]-V-x-[KD]-[FY]-[LIV]-[GHS]-Y-K-L- [SR]-Q-[RK]-G-[HY]-x-[CW].

g) Bromodomain (bromodomain). SEQ ID NOS:4036 and 4489, and thus the corresponding sequences they validate, represent polynucleotides encoding a polypeptide having a bromodomain region (Haynes et al., 1992, Nucleic Acids Res. 20:2693-2603, Tamkun et al., 1992, Cell 68:561-572, and Tamkun, 1995, Curr. Opin. Genet. Dev. 5:473-

OCID: <WO___9938972A2_I_>

477), which is a conserved region of about 70 amino acids found in the following proteins: 1) Higher eukaryotes transcription initiation factor TFIID 250 Kd subunit (TBP-associated factor p250) (gene CCG1); P250 is associated with the TFIID TATA-box binding protein and seems essential for progression of the G1 phase of the cell cycle. 2) Human RING3, a protein of unknown function encoded in the MHC class II locus; 3) Mammalian CREB-binding protein (CBP), which mediates cAMP-gene regulation by binding specifically to phosphorylated CREB protein; 4) Mammalian homologs of brahma, including three brahma-like human: SNF2a(hBRM), SNF2b, and BRG1; 5) Human BS69, a protein that binds to adenovirus E1A and inhibits E1A transactivation; 6) Human peregrin (or Br140).

The bromodomain is thought to be involved in protein-protein interactions and may be important for the assembly or activity of multicomponent complexes involved in transcriptional activation. The consensus pattern, which spans a major part of the bromodomain, is: [STANVF]-x(2)-F-x(4)-[DNS]-x(5,7)-[DENQTF]-Y-[HFY]-x(2)-[LIVMFY]-x(3)-[LIVM]-x(4)-[LIVM]-x(6,8)-Y-x(12,13)-[LIVM]-x(2)-N-[SACF]-x(2)-[FY].

h) Basic Region Plus Leucine Zipper Transcription Factors (BZIP). SEQ ID NO:3408, 2951, and 4850, and thus the corresponding sequences these sequences validate, represent polynucleotides encoding a novel member of the family of basic region plus leucine zipper transcription factors. The bZIP superfamily (Hurst, *Protein Prof.* (1995) *2*:105; and Ellenberger, *Curr. Opin. Struct. Biol.* (1994) *4*:12) of eukaryotic DNA-binding transcription factors encompasses proteins that contain a basic region mediating sequence-specific DNA-binding followed by a leucine zipper required for dimerization. Members of the family include transcription factor AP-1, which binds selectively to enhancer elements in the cis control regions of SV40 and metallothionein IIA. AP-1, also known as c-jun, is the cellular homolog of the avian sarcoma virus 17 (ASV17) oncogene v-jun.
Other members of this protein family include jun-B and jun-D, probable transcription factors that are highly similar to jun/AP-1; the fos protein, a proto-oncogene that forms a non-covalent dimer with c-jun; the fos-related proteins fra-1, and fos B; and mammalian cAMP response element (CRE) binding proteins CREB, CREM, ATF-1, ATF-3, ATF-4, ATF-5, ATF-6 and LRF-1. The consensus pattern for this protein family is: [KR]-x(1,3)-[RKSAQ]-N-x(2)-[SAQ](2)-x-[RKTAENQ]-x-R-x-[RK].

i) Cyclins (cyclin). SEQ ID NOS:3618, 3895, and 4536, and thus the corresponding sequences these sequences validate, represent polynucleotides encoding

cyclins, and SEQ ID NO:55 and 56, respectively, show the corresponding full-length polynucleotides. SEQ ID NO:57 and 58 show, respectively, the translations of SEQ ID NO:55 and 56. Cyclins (Nurse, 1990, Nature 344:503-508; Norbury et al., 1991, Curr. Biol. 1:23-24; and Lew et al., 1992, Trends Cell Biol. 2:77-81) are eukaryotic proteins that play an active role in controlling nuclear cell division cycles. There are two main groups of cyclins. G2/M cyclins are essential for the control of the cell cycle at the G2/M (mitosis) transition. G2/M cyclins accumulate steadily during G2 and are abruptly destroyed as cells exit from mitosis (at the end of the M-phase). G1/S cyclins are essential for the control of the cell cycle at the G1/S (start) transition.

The best conserved region is in the central part of the cyclins' sequences, known as the "cyclin-box," from which a 32 residue consensus pattern was derived: R-x(2)-[LIVMSA]-x(2)-[FYWS]-[LIVM]-x(8)-[LIVMFC]-x(4)-[LIVMFYA]-x(2)-[STAGC]-[LIVMFYQ]-x-[LIVMFYC]-[LIVMFY]-D-[RKH]-[LIVMFYW].

j) Eukaryotic thiol (cysteine) proteases active sites (Cys-protease). SEQ ID NOS:3344, 3684, 3688, and 4801, and thus also the sequences they validate, repreasent polynucleotides encoding proteins having a eukaryotic thiol (cysteine) protease active site. Eukaryotic thiol proteases (Dufour E., *Biochimie* (1988) *70*:1335); are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases that belong to this family are: 1) vertebrate lysosomal cathepsins B (Kirschke H., et al., *Protein Prof.* (1995) *2*:1587-1643); 2) vertebrate lysosomal dipeptidyl peptidase I (also known as cathepsin C) (Kirschke H., et al., *supra*); 3) vertebrate calpains (Calpains are intracellular calcium-activated thiol protease that contain both an N-terminal catalytic domain and a C-terminal calcium-binding domain); 4) mammalian cathepsin K, which seems involved in osteoclastic bone resorption (Shi G.-P., et al., *FEBS Lett.* (1995) *357*:129); 5) human cathepsin O ([4] Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. *J. Biol. Chem.* (1994) *269*:27136); 6) bleomycin hydrolase (which catalyzes the inactivation of the antitumor drug BLM (a glycopeptide)); 7) Plant enzymes such as: barley aleurain, EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin; papaya latex papin, chymopapain, caricain, and proteinase IV; pea turgor-responsive protein 15A; pineapple stem bromelain; rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, Arabidopsis thaliana A494, RD19A and RD21A; 8) - House-dust

DOCID: <WO__9938972A2_I_>

mites allergens DerP1 and EurM1; 9) cathepsin B-like proteinases from the worms Caenorhabditis elegans (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), Schistosoma mansoni (antigen SM31) and Japonica (antigen SJ31), Haemonchus contortus (genes AC-1 and AC-2), and Ostertagia ostertagi (CP-1 and CP-3); 10) slime mold cysteine proteinases CP1 and CP2; 11) cruzipain from Trypanosoma cruzi and brucei; 12) throphozoite cysteine proteinase (TCP) from various Plasmodium species; 13) proteases from Leishmania mexicana, Theileria annulata and Theileria parva; 14) Baculoviruses cathepsin-like enzyme (v-cath); 15) Drosophila small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain; 16) yeast thiol protease BLH1/YCP1/LAP3; 17) Caenorhabditis elegans hypothetical protein C06G4.2, a calpain-like protein.

In addition, two bacterial peptidases are also part of this family: 1) aminopeptidase C from Lactococcus lactis (gene pepC) (Chapot-Chartier M.P., et al., *Appl. Environ. Microbiol.* (1993) *59*:330); and 2) thiol protease tpr from Porphyromonas gingivalis. Three other proteins are structurally related to this family, but may have lost their proteolytic activity. These include: 1) soybean oil body protein P34 (which has its active site cysteine replaced by a glycine); 2) rat testin (which is a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine); and 3) Plasmodium falciparum serine-repeat protein (SERA) (which is the major blood stage antigen and possesses a C-terminal thiol-protease-like domain (Higgins D.G., et al., *Nature* (1989) *340*:604), with the active site cysteine is replaced by a serine).

The sequences around the three active site residues are well conserved and can be used as signature patterns:

Consensus pattern #1: Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV] (where C is the active site residue)

Consensus pattern #2: [LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH] (where H is the active site residue);

Consensus patern #3: [FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G- [LFYW]-[LIVMFYG]-x-[LIVMF] (where N is the active site residue).

k) Phorbol Esters/Diacylglycerol Binding (DAG_PE_bind). SEQ ID NO:4659, and thus the sequence it validates, represents a polynucleotide encoding a protein belonging to the family including phorbol esters/diacylglycerol binding proteins. Diacylglycerol (DAG) is an important second messenger. Phorbol esters (PE) are analogues of DAG and potent tumor promoters that cause a variety of physiological changes when administered to both

cells and tissues. DAG activates a family of serine/threonine protein kinases, collectively known as protein kinase C (PKC) (Azzi *et al.*, *Eur. J. Biochem.* (1992) *208*:547). Phorbol esters can directly stimulate PKC. The N-terminal region of PKC, known as C1, has been shown (Ono *et al.*, *Proc. Natl. Acad. Sci. USA* (1989) *86*:4868) to bind PE and DAG in a phospholipid and zinc-dependent fashion. The C1 region contains one or two copies (depending on the isozyme of PKC) of a cysteine-rich domain about 50 amino-acid residues long and essential for DAG/PE-binding. Such a domain has also been found in, for example, the following proteins.

(1) Diacylglycerol kinase (EC 2.7.1.107) (DGK) (Sakane *et al.*, *Nature* (1990) *344*:345), the enzyme that converts DAG into phosphatidate. It contains two copies of the DAG/PE-binding domain in its N-terminal section. At least five different forms of DGK are known in mammals; and

(2) N-chimaerin, a brain specific protein which shows sequence similarities with the BCR protein at its C-terminal part and contains a single copy of the DAG/PE-binding domain at its N-terminal part. It has been shown (Ahmed *et al.*, *Biochem. J.* (1990) *272*:767, and Ahmed *et al.*, *Biochem. J.* (1991) *280*:233) to be able to bind phorbol esters.

The DAG/PE-binding domain binds two zinc ions; the ligands of these metal ions are probably the six cysteines and two histidines that are conserved in this domain. The signature pattern completely spans the DAG/PE domain. The consensus pattern is: H-x-[LIVMFYW]-x(8,11)-C-x(2)-C-x(3)-[LIVMFC]-x(5,10)-C-x(2)-C-x(4)-[HD]-x(2)-C-x(5,9)-C. All the C and H are probably involved in binding zinc.

l) <u>DEAD and DEAH box families ATP-dependent helicases signatures (Dead_box_helic).</u> SEQ ID NOS:4821 and 5083, and thus the sequences they validate, represent polynucleotides encoding a novel member of the DEAD box family. A number of eukaryotic and prokaryotic proteins have been characterized (Schmid S.R., et al., *Mol. Microbiol.* (1992) *6*:283; Linder P., et al., *Nature* (1989) *337*:121; Wassarman D.A., et al., *Nature* (1991) *349*:463) on the basis of their structural similarity. All are involved in ATP-dependent, nucleic-acid unwinding. Proteins currently known to belong to this family are:

1) Initiation factor eIF-4A. Found in eukaryotes, this protein is a subunit of a high molecular weight complex involved in 5'cap recognition and the binding of mRNA to ribosomes. It is an ATP-dependent RNA-helicase.

2) PRP5 and PRP28. These yeast proteins are involved in various ATP-requiring steps of the pre-mRNA splicing process.

)OCID: <WO__9938972A2_I_>

3) PI10, a mouse protein expressed specifically during spermatogenesis.

4) An3, a Xenopus putative RNA helicase, closely related to PI10.

5) SPP81/DED1 and DBP1, two yeast proteins involved in pre-mRNA splicing and related to PI10.

6) Caenorhabditis elegans helicase glh-1.

7) MSS116, a yeast protein required for mitochondrial splicing.

8) SPB4, a yeast protein involved in the maturation of 25S ribosomal RNA.

9) p68, a human nuclear antigen. p68 has ATPase and DNA-helicase activities in vitro. It is involved in cell growth and division.

10) Rm62 (p62), a Drosophila putative RNA helicase related to p68.

11) DBP2, a yeast protein related to p68.

12) DHH1, a yeast protein.

13) DRS1, a yeast protein involved in ribosome assembly.

14) MAK5, a yeast protein involved in maintenance of dsRNA killer plasmid.

15) ROK1, a yeast protein.

16) ste13, a fission yeast protein.

17) Vasa, a Drosophila protein important for oocyte formation and specification of embryonic posterior structures.

18) Me31B, a Drosophila maternally expressed protein of unknown function.

19) dbpA, an Escherichia coli putative RNA helicase.

20) deaD, an Escherichia coli putative RNA helicase which can suppress a mutation in the rpsB gene for ribosomal protein S2.

21) rhlB, an Escherichia coli putative RNA helicase.

22) rhlE, an Escherichia coli putative RNA helicase.

23) rmB, an Escherichia coli protein that shows RNA-dependent ATPase activity, which interacts with 23S ribosomal RNA.

24) Caenorhabditis elegans hypothetical proteins T26G10.1, ZK512.2 and ZK686.2.

25) Yeast hypothetical protein YHR065c.

26) Yeast hypothetical protein YHR169w.

27) Fission yeast hypothetical protein SpAC31A2.07c.

28) Bacillus subtilis hypothetical protein yxiN.

All of the above proteins share a number of conserved sequence motifs. Some of them are specific to this family while others are shared by other ATP-binding proteins or by proteins belonging to the helicases 'superfamily' (Hodgman T.C., *Nature* (1988) *333*:22 and *Nature* (1988) *333*:578 (Errata); http://www.expasy.ch/www/linder/HELICASES_TEXT.html). One of these motifs, called the 'D-E-A-D-box', represents a special version of the B motif of ATP-binding proteins. Some other proteins belong to a subfamily which have His instead of the second Asp and are thus said to be 'D-E-A-H-box' proteins (Wassarman D.A., et al., *Nature* (1991) *349*:463; Harosh I., et al., *Nucleic Acids Res.* (1991) *19*:6331; Koonin E.V., et al., *J. Gen. Virol.* (1992) *73*:989; http://www.expasy.ch/www/linder/HELICASES_TEXT.html). Proteins currently known to belong to this DEAH subfamily are:

1) PRP2, PRP16, PRP22 and PRP43. These yeast proteins are all involved in various ATP-requiring steps of the pre-mRNA splicing process. 2) Fission yeast prh1, which my be involved in pre-mRNA splicing. 3) Male-less (mle), a Drosophila protein required in males, for dosage compensation of X chromosome linked genes. 4) RAD3 from yeast. RAD3 is a DNA helicase involved in excision repair of DNA damaged by UV light, bulky adducts or cross-linking agents. Fission yeast rad15 (rhp3) and mammalian DNA excision repair protein XPD (ERCC-2) are the homologs of RAD3. 5) Yeast CHL1 (or CTF1), which is important for chromosome transmission and normal cell cycle progression in G(2)/M. 6) Yeast TPS1. 7) Yeast hypothetical protein YKL078w. 8) Caenorhabditis elegans hypothetical proteins C06E1.10 and K03H1.2. 9) Poxviruses' early transcription factor 70 Kd subunit which acts with RNA polymerase to initiate transcription from early gene promoters. 10) I8, a putative vaccinia virus helicase. 11) hrpA, an Escherichia coli putative RNA helicase.

The following signature patterns are used to identify member for both subfamilies:

Consensus pattern: [LIVMF](2)-D-E-A-D-[RKEN]-x-[LIVMFYGSTN]

Consensus pattern: [GSAH]-x-[LIVMF](3)-D-E-[ALIV]-H-[NECR].

m) EF Hand (EFhand). Several of the validation sequences, and thus the sequences they validate, correspond to polynucleotides encoding a novel protein in the family of EF-hand proteins. Many calcium-binding proteins belong to the same evolutionary family and share a type of calcium-binding domain known as the EF-hand (Kawasaki *et al., Protein. Prof.* (1995) *2*:305-490). This type of domain consists of a twelve residue loop flanked on both sides by a twelve residue alpha-helical domain. In an EF-hand loop the calcium ion is

)OCID: <WO___9938972A2_I_>

coordinated in a pentagonal bipyramidal configuration. The six residues involved in the binding are in positions 1, 3, 5, 7, 9 and 12; these residues are denoted by X, Y, Z, -Y, -X and -Z. The invariant Glu or Asp at position 12 provides two oxygens for liganding Ca (bidentate ligand).

Proteins known to contain EF-hand regions include: Calmodulin (Ca=4, except in yeast where Ca=3) ("Ca=" indicates approximate number of EF-hand regions); diacylglycerol kinase (EC 2.7.1.107) (DGK) (Ca=2); 2) FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) from mammals (Ca=1); guanylate cyclase activating protein (GCAP) (Ca=3); MIF related proteins 8 (MRP-8 or CFAG) and 14 (MRP-14) (Ca=2); myosin regulatory light chains (Ca=1); oncomodulin (Ca=2); osteonectin (basement membrane protein BM-40) (SPARC); and proteins that contain an "osteonectin" domain (QR1, matrix glycoprotein SC1).

The consensus pattern includes the complete EF-hand loop as well as the first residue which follows the loop and which seem to always be hydrophobic: D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].

n) Ets Domain (Ets_Nterm). SEQ ID NO:2849, and thus the sequence it validates, represents a polynucleotide encoding a polypeptide with N-terminal homology in ETS domain. Proteins of this family contain a conserved domain, the "ETS-domain," that is involved in DNA binding. The domain appears to recognize purine-rich sequences; it is about 85 to 90 amino acids in length, and is rich in aromatic and positively charged residues (Wasylyk, et al., , *Eur. J. Biochem.* (1993) *211*:718).

The *ets* gene family encodes a novel class of DNA-binding proteins, each of which binds a specific DNA sequence. These proteins comprise an *ets* domain that specifically interacts with sequences containing the common core tri-nucleotide sequence GGA. In addition to an *ets* domain, native *ets* proteins comprise other sequences which can modulate the biological specificity of the protein. *Ets* genes and proteins are involved in a variety of essential biological processes including cell growth, differentiation and development, and three members are implicated in oncogenic process.

o) Type II fibronectin collagen-binding domain (FntypeII). A few of the validation sequences, and thus the sequences they validate, represent polynucleotides encoding a polypeptide having a type II fibronectin collagen binding domain. Fibronectin is a plasma protein that binds cell surfaces and various compounds including collagen, fibrin, heparin,

DNA, and actin. The major part of the sequence of fibronectin consists of the repetition of three types of domains, which are called type I, II, and III (Skorstengaard K., et al., *Eur. J. Biochem.* (1986) *161*:441). Type II domain is approximately forty residues long, contains four conserved cysteines involved in disulfide bonds and is part of the collagen-binding region of fibronectin. In fibronectin the type II domain is duplicated. Type II domains have also been found in the following proteins: 1) blood coagulation factor XII (Hageman factor) (1 copy); 2) bovine seminal plasma proteins PDC-109 (BSP-A1/A2) and BSP-A3 (Seidah N.G., et al., *Biochem. J.* (1987) *243*:195. (twice); 3) cation-independent mannose-6-phosphate receptor (which is also the insulin-like growth factor II receptor) Kornfeld S., *Annu. Rev. Biochem.* (1992) *61*:307) (1 copy); 4) Mannose receptor of macrophages (Taylor M.E., et al., *J. Biol. Chem.* (1990) *265*:12156) (1 copy); 5) 180 Kd secretory phospholipase A2 receptor (1 copy) Lambeau G., et al., *J. Biol. Chem.* (1994) *269*:1575; 6) DEC-205 receptor (1 copy); 6) Jiang W., et al., *Nature* (1995) *375*:151); 7) 72 Kd type IV collagenase (EC 3.4.24.24) (MMP-2) (Collier I.E., et al., *J. Biol. Chem.* (1988) *263*:6579) (3 copies); 7) 92 Kd type IV collagenase (EC 3.4.24.24) (MMP-9) (3 copies); 8) Hepatocyte growth factor activator (Miyazawa K., et al., *J. Biol. Chem.* (1993) *268*:10024) (1 copy).

A schematic representation of the position of the invariant residues and the topology of the disulfide bonds in fibronectin type II domain is shown below:

xxCxxPFx#xxxxxxxCxxxxxxxxWCxxxxx#xxx#x#Cxx

where 'C' represents the conserved cysteine involved in a disulfide bond and '#' represents a large hydrophobic residue. The consensus pattern for identifying members of this family, which pattern spans this entire domain, is: C-x(2)-P-F-x-[FYWI]-x(7)-C-x(8,10)-W-C-x(4)-[DNSR]-[FYW]- x(3,5)-[FYW]-x-[FYWI]-C (where the four C's are involved in disulfide bonds).

p) G-Protein Alpha Subunit (G-alpha). Several of the validation sequences, and thus the sequences they validate, correspond to a gene encoding a novel polypeptide of the G-protein alpha subunit family. Guanine nucleotide binding proteins (G-proteins) are a family of membrane-associated proteins that couple extracellularly-activated integral-membrane receptors to intracellular effectors, such as ion channels and enzymes that vary the concentration of second messenger molecules. G-proteins are composed of 3 subunits (alpha, beta and gamma) which, in the resting state, associate as a trimer at the inner face of

OCID: <WO__9938972A2_I_>

the plasma membrane. The alpha subunit has a molecule of guanosine diphosphate (GDP) bound to it. Stimulation of the G-protein by an activated receptor leads to its exchange for GTP (guanosine triphosphate). This results in the separation of the alpha from the beta and gamma subunits, which always remain tightly associated as a dimer. Both the alpha and beta-gamma subunits are then able to interact with effectors, either individually or in a cooperative manner. The intrinsic GTPase activity of the alpha subunit hydrolyses the bound GTP to GDP. This returns the alpha subunit to its inactive conformation and allows it to reassociate with the beta-gamma subunit, thus restoring the system to its resting state.

G-protein alpha subunits are 350-400 amino acids in length and have molecular weights in the range 40-45 kDa. Seventeen distinct types of alpha subunit have been identified in mammals. These fall into 4 main groups on the basis of both sequence similarity and function: alpha-s, alpha-q, alpha-i and alpha-12 (Simon *et al.*, *Science* (1993) *252*:802). Many alpha subunits are substrates for ADP-ribosylation by cholera or pertussis toxins. They are often N-terminally acylated, usually with myristate and/or palmitoylate, and these fatty acid modifications are probably important for membrane association and high- affinity interactions with other proteins. The atomic structure of the alpha subunit of the G-protein involved in mammalian vision, transducin, has been elucidated in both GTP- and GDB-bound forms, and shows considerable similarity in both primary and tertiary structure in the nucleotide-binding regions to other guanine nucleotide binding proteins, such as p21-ras and EF-Tu.

q) Helicases conserved C-terminal domain (helicase_C). SEQ ID NOS:2503, 4469, and 5020, and thus the sequences they validate, represent polynucleotides encoding novel members of the DEAD/H helicase family. The DEAD and DEAH families are described above.

r) Homeobox domain (homeobox). SEQ ID NO:4241, and thus the sequence it validates, represents a polynucleotide encoding a protein having a homeobox domain. The 'homeobox' is a protein domain of 60 amino acids (Gehring In: Guidebook to the Homebox Genes, Duboule D., Ed., pp1-10, Oxford University Press, Oxford, (1994); Buerglin In: Guidebook to the Homebox Genes, pp25-72, Oxford University Press, Oxford, (1994); Gehring *Trends Biochem. Sci.* (1992) *17*:277-280; Gehring *et alAnnu. Rev. Genet.* (1986) *20*:147-173; Schofield *Trends Neurosci.* (1987) *10*:3-6; http://copan.bioz.unibas.ch/homeo.html) first identified in number of Drosophila homeotic and segmentation proteins. It is extremely well conserved in many other animals, including vertebrates. This domain

binds DNA through a helix-turn-helix type of structure. Several proteins that contain a homeobox domain play an important role in development. Most of these proteins are sequence-specific DNA-binding transcription factors. The homeobox domain is also very similar to a region of the yeast mating type proteins. These are sequence-specific DNA-binding proteins that act as master switches in yeast differentiation by controlling gene expression in a cell type-specific fashion.

A schematic representation of the homeobox domain is shown below. The helix-turn-helix region is shown by the symbols 'H' (for helix), and 't' (for turn).

```
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHHHttttHHHHHHHHHHxxxxxxxxxx
1                                                                60
```

The pattern detects homeobox sequences 24 residues long and spans positions 34 to 57 of the homeobox domain. The consensus pattern is as follows: [LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)-[RKNAIMW].

x) MAP kinase kinase (mkk). Several validation sequences, and thus the sequences they validate, represent novel members of the MAP kinase kinase family. MAP kinases (MAPK) are involved in signal transduction, and are important in cell cycle and cell growth controls. The MAP kinase kinases (MAPKK) are dual-specificity protein kinases which phosphorylate and activate MAP kinases. MAPKK homologues have been found in yeast, invertebrates, amphibians, and mammals. Moreover, the MAPKK/MAPK phosphorylation switch constitutes a basic module activated in distinct pathways in yeast and in vertebrates. MAPKK regulation studies have led to the discovery of at least four MAPKK convergent pathways in higher organisms. One of these is similar to the yeast pheromone response pathway which includes the ste11 protein kinase. Two other pathways require the activation of either one or both of the serine/threonine kinase-encoded oncogenes c-Raf-1 and c-Mos. Additionally, several studies suggest a possible effect of the cell cycle control regulator cyclin-dependent kinase 1 (cdc2) on MAPKK activity. Finally, MAPKKs are apparently essential transducers through which signals must pass before reaching the nucleus. For review, see, *e.g.*, Biologique *Biol Cell* (1993) *79*:193-207; Nishida *et al.*, *Trends Biochem Sci* (1993) *18*:128-31; Ruderman *Curr Opin Cell Biol* (1993) *5*:207-13; Dhanasekaran *et al.*, *Oncogene* (1998) *17*:1447-55; Kiefer *et al.*, *Biochem Soc Trans* (1997) *25*:491-8; and Hill, *Cell Signal* (1996) *8*:533-44.

y) 3'5'-cyclic nucleotide phosphodiesterases signature (PDEase). SEQ ID NO:4482, and thus the sequence it validates, represents a polynucleotide encoding a novel 3'5'-cyclic nucleotide phosphodiesterases (PDEases). PDEases catalyze the hydrolysis of cAMP or cGMP to the corresponding nucleoside 5' monophosphates (Charbonneau H., et al, *Proc. Natl. Acad. Sci. U.S.A.* (1986) *83*:9308). There are at least seven different subfamilies of PDEases (Beavo J.A., et al., *Trends Pharmacol. Sci.* (1990) *11*:150; http://weber.u.washington.edu/~pde/: 1) Type 1, calmodulin/calcium-dependent PDEases; 2) Type 2, cGMP-stimulated PDEases; 3) Type 3, cGMP-inhibited PDEases; 4) Type 4, cAMP-specific PDEases.; 5) Type 5, cGMP-specific PDEases; 6) Type 6, rhodopsin-sensitive cGMP-specific PDEases; and 7) Type 7, High affinity cAMP-specific PDEases.

All PDEase forms share a conserved domain of about 270 residues. The signature pattern is determined from a stretch of 12 residues that contains two conserved histidines: H-D-[LIVMFY]-x-H-x-[AG]-x(2)-[NQ]-x-[LIVMFY].

z) Protein Kinase (protkinase). Several validation sequences, and thus the sequences they validate, represent polynucleotides encoding protein kinases. Protein kinases catalyze phosphorylation of proteins in a variety of pathways, and are implicated in cancer. Eukaryotic protein kinases (Hanks S.K., *et al.*, *FASEB J.* (1995) *9*:576; Hunter T., *Meth. Enzymol.* (1991) *200*:3; Hanks S.K., *et al.*, *Meth. Enzymol.* (1991) *200*:38; Hanks S.K., *Curr. Opin. Struct. Biol.* (1991) *1*:369; Hanks S.K., *et al.*, *Science* (1988) *241*:42) are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. Two of the conserved regions are the basis for the signature pattern in the protein kinase profile. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. The second region, which is located in the central part of the catalytic domain, contains a conserved aspartic acid residue which is important for the catalytic activity of the enzyme (Knighton D.R., *et al.*, *Science* (1991) *253*:407). The protein kinase profile includes two signature patterns for this second region: one specific for serine/threonine kinases and the other for tyrosine kinases. A third profile is based on the alignment in (Hanks S.K., *et al.*, *FASEB J.* (1995) *9*:576) and covers the entire catalytic domain. The consensus patterns are as follows:

OCID: <WO__9938972A2_I_>

1) Consensus pattern: [LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K, where K binds ATP. The majority of known protein kinases are detected by this pattern. Proteins kinases that are not detected by this consensus include viral kinases, which are quite divergent in this region and are completely missed bythis pattern.

2) Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3), where D is an active site residue. This consensus sequence identifies most serine/threonine-specific protein kinases with only 10 exceptions. Half of the exceptions are viral kinases, while the other exceptions include Epstein-Barr virus BGLF4 and Drosophila ninaC, which have Ser and Arg, respectively, instead of the conserved Lys. These latter two protein kinases are detected by the tyrosine kinase specific pattern described below.

3) Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC], where D is an active site residue. All tyrosine-specific protein kinases are detected by this consensus pattern, with the exception of human ERBB3 and mouse blk. This pattern also detects most bacterial aminoglycoside phosphotransferases (Benner S., *Nature* (1987) *329*:21; Kirby R., *J. Mol. Evol.* (1992) *30*:489) and herpesviruses ganciclovir kinases (Littler E., *et al.*, *Nature* (1992) *358*:160), which are structurally and evolutionary related to protein kinases.

The protein kinase profile also detects receptor guanylate cyclases and 2-5A-dependent ribonucleases. Sequence similarities between these two families and the eukaryotic protein kinase family have been noticed previously. The profile also detects Arabidopsis thaliana kinase-like protein TMKL1 which seems to have lost its catalytic activity.

If a protein analyzed includes the two of the above protein kinase signatures, the probability of it being a protein kinase is close to 100%. Eukaryotic-type protein kinases have also been found in prokaryotes such as Myxococcus xanthus (Munoz-Dorado J., *et al.*, *Cell* (1991) *67*:995) and Yersinia pseudotuberculosis. The patterns shown above has been updated since their publication in (Bairoch A., *et al.*, *Nature* (1988) *331*:22).

aa) Ras family proteins (ras). SEQ IDNO:3671, and thus the sequence it validates, represent polynucleotides encoding the ras family of small GTP/GDP-binding proteins (Valencia et al., 1991, Biochemistry 30:4637-4648). Ras family members generally require

a specific guanine nucleotide exchange factor (GEF) and a specific GTPase activating protein (GAP) as stimulators of overall GTPase activity. Among ras-related proteins, the highest degree of sequence conservation is found in four regions that are directly involved in guanine nucleotide binding. The first two constitute most of the phosphate and Mg2+ binding site (PM site) and are located in the first half of the G-domain. The other two regions are involved in guanosine binding and are located in the C-terminal half of the molecule. Motifs and conserved structural features of the ras-related proteins are described in Valencia et al., 1991, Biochemistry 30:4637-4648.

A major consensus pattern of ras proteins is: D-T-A-G-Q-E-K-[LF]-G-G-L-R-[DE]-G-Y-Y.

bb) Thioredoxin family active site (Thioredox). SEQ ID NO:3936, and thus the sequence it validates, represent a polynucleotide encoding a protein having a thioredoxin family active site. Thioredoxins (Holmgren A., *Annu. Rev. Biochem.* (1985) *54*:237; Gleason F.K., et al., *FEMS Microbiol. Rev.* (1988) *54*:271; Holmgren A. *J. Biol. Chem.* (1989) *264*:13963; Eklund H., et al. *Proteins* (1991) *11*:13) are small proteins of approximately one hundred amino- acid residues which participate in various redox reactions via the reversible oxidation of an active center disulfide bond. They exist in either a reduced form or an oxidized form where the two cysteine residues are linked in an intramolecular disulfide bond. Thioredoxin is present in prokaryotes and eukaryotes and the sequence around the redox-active disulfide bond is well conserved.

A number of eukaryotic proteins contain domains evolutionary related to thioredoxin, and all of them are protein disulphide isomerases (PDI). PDI (Freedman R.B., et al., *Biochem. Soc. Trans.* (1988) *16*:96; Kivirikko K.I., et al., *FASEB J.* (1989) *3*:1609; Freedman R.B., et al. *Trends Biochem. Sci.* (1994) *19*:331) is an endoplasmic reticulum enzyme that catalyzes the rearrangement of disulfide bonds in various proteins. The various forms of PDI which are currently known are: 1) PDI major isozyme; a multifunctional protein that also function as the beta subunit of prolyl 4-hydroxylase (EC 1.14.11.2), as a component of oligosaccharyl transferase (EC 2.4.1.119), as thyroxine deiodinase, as glutathione-insulin transhydrogenase, and as a thyroid hormone-binding protein; 2) ERp60 (ER-60; 58 Kd microsomal protein), which is a protease; 3) ERp72; and 4) P5.

All PDI contains two or three (ERp72) copies of the thioredoxin domain. The consensus pattern is: [LIVMF]-[LIVMSTA]-x-[LIVMFYC]-[FYWSTHE]-x(2)-

OCID: <WO__9938972A2_I_>

[FYWGTN]-C- [GATPLVE]-[PHYWSTA]-C-x(6)-[LIVMFYWT] (where the two C's form the redox-active bond.

cc) TNFR/NGFR family cysteine-rich region (TNFR_c6). SEQ ID NO:3927, and thus the sequence it validates, represent a polynucleotide encoding a protein having a TNFR/NGFR family cysteine-rich region. A number of proteins, some of which are known to be receptors for growth factors, have been found to contain a cysteine-rich domain of about 110 to 160 amino acids in their N-terminal part, that can be subdivided into four (or in some cases, three) modules of about 40 residues containing 6 conserved cysteines. Proteins known to belong to this family (Mallet S., et al., *Immunol. Today* (1991) 12:220; Sprang S.R., *Trends Biochem. Sci.* (1990) 15:366; Krammer P.H., et al., *Curr. Biol.* (1992) 2:383; Bazan J.F., *Curr. Biol.* (1993) *3*:603) are: 1) Tumor Necrosis Factor type I and type II receptors (TNFR) (Both receptors bind TNF-alpha and TNF-beta, but are only similar in the cysteine-rich region.); 2) Shope fibroma virus soluble TNF receptor (protein T2); 3) Lymphotoxin alpha/beta receptor; 4) Low-affinity nerve growth factor receptor (LA-NGFR); 5) CD40 (Bp50), the receptor for the CD40L (or TRAP) cytokine; 6) CD27, the receptor for the CD27L cytokine; 8) CD30, the receptor for the CD30L cytokine; 9) T-cell protein 4-1BB, the receptor for the 4-1BBL putative cytokine; 10) FAS antigen (or APO-1), the receptor for FASL, a protein involved in apoptosis (programmed cell death); 11) T-cell antigen OX40, the receptor for the OX40L cytokine; 12) Wsl-1, a receptor (for a yet undefined ligand) that mediates apoptosis; 13) Vaccinia virus protein A53 (SalF19R).

The six cysteines all involved in intrachain disulfide bonds (Banner D.W., et al, *Cell* (1993) *73*:431). A schematic representation of the structure of the 40 residue module of these receptors is shown below:

xCxxxxxxxxxxxxxCxCxxCxxxxxxxxxCxxxxCxx

where 'C' represents the conserved cysteine involved in a disulfide bond. The signature pattern for the cysteine-rich region is based mainly on the position of the six conserved cysteines in each of the repeats: Consensus pattern: C-x(4,6)-[FYH]-x(5,10)-C-x(0,2)-C-x(2,3)-C-x(7,11)-C-x(4,6)-[DNEQSKP]-x(2)-C (where the six C's are involved in disulfide bonds).

dd) Four Transmembrane Integral Membrane Proteins (transmembrane4). Several of the validation sequences, and thus the sequences they validate, correspond to a sequence encoding a polypeptide that is a member of the 4 transmembrane segments integral

membrane protein family (transmembrane 4 family). The transmembrane 4 family of proteins includes a number of evolutionarily-related eukaryotic cell surface antigens (Levy *et al.*, *J. Biol. Chem.*, (1991) *266*:14597; Tomlinson *et al.*, *Eur. J. Immunol.* (1993) *23*:136; Barclay *et al.* The leucocyte antigen factbooks. (1993) Academic Press, London/San Diego). The proteins belonging to this family include: 1) Mammalian antigen CD9 (MIC3), which is involved in platelet activation and aggregation; 2) Mammalian leukocyte antigen CD37, expressed on B lymphocytes; 3) Mammalian leukocyte antigen CD53 (OX-44), which is implicated in growth regulation in hematopoietic cells; 4) Mammalian lysosomal membrane protein CD63 (melanoma-associated antigen ME491; antigen AD1); 5) Mammalian antigen CD81 (cell surface protein TAPA-1), which is implicated in regulation of lymphoma cell growth; 6) Mammalian antigen CD82 (protein R2; antigen C33; Kangai 1 (KAI1)), which associates with CD4 or CD8 and delivers costimulatory signals for the TCR/CD3 pathway; 7) Mammalian antigen CD151 (SFA-1; platelet-endothelial tetraspan antigen 3 (PETA-3)); 8) Mammalian cell surface glycoprotein A15 (TALLA-1; MXS1); 9) Mammalian novel antigen 2 (NAG-2); 10) Human tumor-associated antigen CO-029; 11) Schistosoma mansoni and japonicum 23 Kd surface antigen (SM23 / SJ23).

The members of the 4 transmembrane family share several characteristics. First, they all are apparently type III membrane proteins, which are integral membrane proteins containing an N-terminal membrane-anchoring domain which is not cleaved during biosynthesis and which functions both as a translocation signal and as a membrane anchor. The family members also contain three additional transmembrane regions, at least seven conserved cysteines residues, and are of approximately the same size (218 to 284 residues). These proteins are collectively know as the "transmembrane 4 superfamily" (TM4) because they span plasma membrane four times.

DOCID: <WO__9938972A2_I_>

A schematic diagram of the domain structure of these proteins is as follows:

```
+-+----+------+---+----+----+-----------------------+----+---+
|| TMa | Extra | TM2| Cyt | TM3 | Extracellular        | TM4 | Cyt|
+-+----+------+---C----C----+------CC-------C----C---+----C---+
             *********
```

where Cyt is the cytoplasmic domain, TMa is the transmembrane anchor; TM2 to TM4 represents transmembrane regions 2 to 4, 'C' are conserved cysteines, and '*'indicates the position of the consensus pattern. The consensus pattern spans a conserved region including two cysteines located in a short cytoplasmic loop between two transmembrane domains: Consensus pattern: G-x(3)-[LIVMF]-x(2)-[GSA]-[LIVMF](2)-G-C-x-[GA]-[STA]- x(2)-[EG]-x(2)-[CWN]-[LIVM](2).

ee) Trypsin (trypsin). SEQ ID NOS:3381, 4684, and 4688, and thus the sequences they validate, correspond to novel serine proteases of the trypsin family. The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases (Brenner S., *Nature* (1988) *334*:528). Proteases known to belong to the trypsin family include: 1) Acrosin; 2) Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C; 3) Cathepsin G; 4) Chymotrypsins; 5) Complement components C1r, C1s, C2, and complement factors B, D and I; 6) Complement-activating component of RA-reactive factor; 7) Cytotoxic cell proteases (granzymes A to H); 8) Duodenase I; 9) Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin).; 10) Enterokinase (EC 3.4.21.9) (enteropeptidase); 11) Hepatocyte growth factor activator; 12) Hepsin; 13) Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin); 14) Plasma kallikrein; 15) Mast cell proteases (MCP) 1 (chymase) to 8; 16) Myeloblastin (proteinase 3) (Wegener's autoantigen); 17) Plasminogen activators (urokinase-type, and tissue-type); 18) Trypsins I, II, III, and IV; 19) Tryptases; 20) Snake venom proteases such as ancrod, batroxobin, cerastobin, flavoxobin, and protein C activator; 21) Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab; 22) Apolipoprotein(a); 23) Blood fluke cercarial protease; 24) Drosophila trypsin like proteases: alpha, easter, snake-locus; 25) Drosophila protease stubble (gene sb); and 26) Major mite fecal allergen Der p

III. All the above proteins belong to family S1 in the classification of peptidases (Rawlings N.D., *et al.*, *Meth. Enzymol.* (1994) *244*:19; http://www.expasy.ch/cgi-bin/lists?peptidas.txt) and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar enough in the regions of the active site residues that they can be picked up by the same patterns.

The consensus patterns for this trypsin protein family are: 1) [LIVM]-[ST]-A-[STAG]-H-C, where H is the active site residue. All sequences known to belong to this class detected by the pattern, except for complement components C1r and C1s, pig plasminogen, bovine protein C, rodent urokinase, ancrod, gyroxin and two insect trypsins; 2) [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]- [LIVMFYWH]-[LIVMFYSTANQH], where S is the active site residue. All sequences known to belong to this family are detected by the above consensus sequences, except for 18 different proteases which have lost the first conserved glycine. If a protein includes both the serine and the histidine active site signatures, the probability of it being a trypsin family serine protease is 100%.

ff) WD Domain, G-Beta Repeats (WD_domain). A few of the validation sequences, and the sequences they validate, represent novel members of the WD domain/G-beta repeat family. Beta-transducin (G-beta) is one of the three subunits (alpha, beta, and gamma) of the guanine nucleotide-binding proteins (G proteins) which act as intermediaries in the transduction of signals generated by transmembrane receptors (Gilman, *Annu. Rev. Biochem.* (1987) *56*:615). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

In higher eukaryotes, G-beta exists as a small multigene family of highly conserved proteins of about 340 amino acid residues. Structurally, G-beta consists of eight tandem repeats of about 40 residues, each containing a central Trp-Asp motif (this type of repeat is sometimes called a WD-40 repeat). Such a repetitive segment has been shown to exist in a number of other proteins including: human LIS1, a neuronal protein involved in type-1 lissencephaly; and mammalian coatomer beta' subunit (beta'-COP), a component of a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport.

The consensus pattern for the WD domain/G-Beta repeat family is: [LIVMSTAC]-

DOCID: <WO___9938972A2_I_>

[LIVMFYWSTAGC]-[LIMSTAG]-[LIVMSTAGC]-x(2)-[DN]-x(2)-[LIVMWSTAC]-x-[LIVMFSTAG]-W-[DEN]-[LIVMFSTAGCN].

gg) wnt Family of Developmental Signaling Proteins (Wnt_dev_sign). Several of the validation sequences, and thus the sequences they validate, correspond to novel members of the wnt family of developmental signaling proteins. Wnt-1 (previously known as int-1), the seminal member of this family, (Nusse R., *Trends Genet.* (1988) *4*:291) is a proto-oncogene induced by the integration of the mouse mammary tumor virus. It is thought to play a role in intercellular communication and seems to be a signalling molecule important in the development of the central nervous system (CNS). The sequence of wnt-1 is highly conserved in mammals, fish, and amphibians. Wnt-1 was found to be a member of a large family of related proteins (Nusse R., *et al.*, *Cell* (1992) *69*:1073; McMahon A.P., *Trends Genet.* (1992) *8*:1; Moon R.T., *BioEssays* (1993) *15*:91) that are all thought to be developmental regulators. These proteins are known as wnt-2 (also known as irp), wnt-3, -3A, -4, -5A, -5B, -6, -7A, -7B, -8, -8B, -9 and -10. At least four members of this family are present in Drosophila; one of them, wingless (wg), is implicated in segmentation polarity.

All these proteins share the following features characteristics of secretory proteins: a signal peptide, several potential N-glycosylation sites and 22 conserved cysteines that are probably involved in disulfide bonds. The Wnt proteins seem to adhere to the plasma membrane of the secreting cells and are therefore likely to signal over only few cell diameters. The consensus pattern, which is based upon a highly conserved region including three cysteines, is as follows: C-K-C-H-G-[LIVMT]-S-G-x-C. All sequences known to belong to this family are detected by the provided consensus pattern.

hh) Protein Tyrosine Phosphatase (Y_phosphatase). Several of the validation sequences, and thus the sequences they validate, represent a polynucleotide encoding a protein tyrosine kinase. Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) (Fischer *et al.*, *Science* (1991) *253*:401; Charbonneau *et al.*, *Annu. Rev. Cell Biol.* (1992) *8*:463; Trowbridge, *J. Biol. Chem.* (1991) *266*:23517; Tonks *et al.*, *Trends Biochem. Sci.* (1989) *14*:497; and Hunter, *Cell* (1989) *58*:1013) catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s).

Soluble PTPases include PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1-like domain and could act at junctions between the membrane and cytoskeleton; PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes that contain two copies of the SH2 domain at its N-terminal extremity.

Dual specificity PTPases include DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1) which dephosphorylates MAP kinase on both Thr-183 and Tyr-185; and DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues.

Structurally, all known receptor PTPases are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPAse domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not.

PTPase domains consist of about 300 amino acids. There are two conserved cysteines and the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important. The consensus pattern for PTPases is: [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY]; C is the active site residue.

ii)Zinc Finger, C2H2 Type (Zincfing_C2H2). Several of the validation sequences, and thus the sequences they validate, correspond to polynucleotides encoding novel members of the of the C2H2 type zinc finger protein family. Zinc finger domains (Klug *et al.*, *Trends Biochem. Sci.* (1987) *12*:464; Evans *et al.*, *Cell* (1988) *52*:1; Payre *et al.*, *FEBS Lett.* (1988) *234*:245; Miller *et al.*, *EMBO J.* (1985) *4*:1609; and Berg, *Proc. Natl. Acad. Sci. USA* (1988) *85*:99) are nucleic acid-binding protein structures first identified in the Xenopus transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger domain is composed of 25 to 30 amino acid residues. Two cysteine or histidine residues are positioned at both extremities of the domain, which are involved in the tetrahedral coordination of a zinc atom. It has been proposed that such a domain interacts with about five nucleotides.

Many classes of zinc fingers are characterized according to the number and

XCID: <WO___9938972A2_I_>

positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines. A number of experimental reports have demonstrated the zinc-dependent DNA or RNA binding property of some members of this class.

Mammalian proteins having a C2H2 zipper include (number in parenthesis indicates number of zinc finger regions in the protein): basonuclin (6), BCL-6/LAZ-3 (6), erythroid krueppel-like transcription factor (3), transcription factors Sp1 (3), Sp2 (3), Sp3 (3) and Sp(4) 3, transcriptional repressor YY1 (4), Wilms' tumor protein (4), EGR1/Krox24 (3), EGR2/Krox20 (3), EGR3/Pilot (3), EGR4/AT133 (4), Evi-1 (10), GLI1 (5), GLI2 (4+), GLI3 (3+), HIV-EP1/ZNF40 (4), HIV-EP2 (2), KR1 (9+), KR2 (9), KR3 (15+), KR4 (14+), KR5 (11+), HF.12 (6+), REX-1 (4), ZfX (13), ZfY (13), Zfp-35 (18), ZNF7 (15), ZNF8 (7), ZNF35 (10), ZNF42/MZF-1 (13), ZNF43 (22), ZNF46/Kup (2), ZNF76 (7), ZNF91 (36), ZNF133 (3).

In addition to the conserved zinc ligand residues, it has been shown that a number of other positions are also important for the structural integrity of the C2H2 zinc fingers. (Rosenfeld *et al., J. Biomol. Struct. Dyn.* (1993) *11*:557) The best conserved position is found four residues after the second cysteine; it is generally an aromatic or aliphatic residue. The consensus pattern for C2H2 zinc fingers is: C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H. The two C's and two H's are zinc ligands.

jj) Zinc finger, C3HC4 type (RING finger), signature (Zincfing_C3H4). SEQ ID NOS:3774 and 4477, and thus the sequences they validate, represent polynucleotides encoding a polypeptide having a C3HC4 type zinc finger signature. A number of eukaryotic and viral proteins contain this signature, which is primarily a conserved cysteine-rich domain of 40 to 60 residues (Borden K.L.B., et al., *Curr. Opin. Struct. Biol.* (1996) *6*:395) that binds two atoms of zinc, and is probably involved in mediating protein-protein interactions. The 3D structure of the zinc ligation system is uniqueto the RING domain and is refered to as the "cross-brace" motif. The spacing of the cysteines in such a domain is C-x(2)-C-x(9 to 39)-C-x(1 to 3)-H-x(2 to 3)-C-x(2)-C-x(4 to 48)-C-x(2)-C. Proteins that include the C3HC4 domain include:

1) Mammalian V(D)J recombination activating protein (RAG1). RAG1 activates the rearrangement of immunoglobulin and T-cell receptor genes.

2) Mouse rpt-1. Rpt-1 is a trans-acting factor that regulates gene expression directed

by the promoter region of the interleukin-2 receptor alpha chainor the LTR promoter region of HIV-1.

3) Human rfp. Rfp is a developmentally regulated protein that may function in male germ cell development. Recombination of the N-terminal section of rfp with a protein tyrosine kinase produces the ret transforming protein.

4) Human 52 Kd Ro/SS-A protein. A protein of unknown function from the Ro/SS-A ribonucleoprotein complex. Sera from patients with systemic lupus erythematosus or primary Sjogren's syndrome often contain antibodies that react with the Ro proteins.

5) Human histocompatibility locus protein RING1.

6) Human PML, a probable transcription factor. Chromosomal translocation ofPML with retinoic receptor alpha creates a fusion protein which is thecause of acute promyelocytic leukemia (APL).

7) Mammalian breast cancer type 1 susceptibility protein (BRCA1) ([E1] http://bioinformatics.weizmann.ac.il/hotmolecbase/entries/brca1.htm).

8) Mammalian cbl proto-oncogene.

9) Mammalian bmi-1 proto-oncogene.

10) Vertebrate CDK-activating kinase (CAK) assembly factor MAT1, a protein that stabilizes the complex between the CDK7 kinase and cyclin H (MAT1 stands for 'Menage A Trois').

11) Mammalian mel-18 protein. Mel-18 which is expressed in a variety of tumorcells is a transcriptional repressor that recognizes and bind a specific DNA sequence.

12) Mammalian peroxisome assembly factor-1 (PAF-1) (PMP35), which is somewhat involved in the biogenesis of peroxisomes. In humans, defects in PAF-1 are responsible for a form of Zellweger syndrome, an autosomal recessivedisorder associated with peroxisomal deficiencies.

13) Human MAT1 protein, which interacts with the CDK7-cyclin H complex.

14) Human RING1 protein.

15) Xenopus XNF7 protein, a probable transcription factor.

16) Trypanosoma protein ESAG-8 (T-LR), which may be involved in the postranscriptional regulation of genes in VSG expression sites or may interact with adenylate cyclase to regulate its activity.

17) Drosophila proteins Posterior Sex Combs (Psc) and Suppressor two of zeste

XOCID: <WO__9938972A2_I_>

(Su(z)2). The two proteins belong to the Polycomb group of genes needed to maintain the segment-specific repression of homeotic selector genes.

18) Drosophila protein male-specific msl-2, a DNA-binding protein which is involved in X chromosome dosage compensation (the elevation of transcription of the male single X chromosome).

19) Arabidopsis thaliana protein COP1 which is involved in the regulation ofphotomorphogenesis.

20) Fungal DNA repair proteins RAD5, RAD16, RAD18 and rad8.

21) Herpesviruses trans-acting transcriptional protein ICP0/IE110. This protein which has been characterized in many different herpesviruses is a trans-activator and/or - repressor of the expression of many viral and cellular promoters.

22) Baculoviruses protein CG30.

23) Baculoviruses major immediate early protein (PE-38).

24) Baculoviruses immediate-early regulatory protein IE-N/IE-2.

25) Caenorhabditis elegans hypothetical proteins F54G8.4, R05D3.4 and T02C1.1.

26) Yeast hypothetical proteins YER116c and YKR017c.

The signature pattern for the C3HC4 finger is based on the central region of the domain:

Consensus pattern: C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA].

Example 4: Differential Expression of Polynucleotides of the Invention: Description of Libraries and Detection of Differential Expression

The relative expression levels of the polynucleotides of the invention was assessed in several libraries prepared from various sources, including cell lines and patient tissue samples. Table 4 provides a summary of these libraries, including the shortened library name (used hereafter), the mRNA source used to prepared the cDNA library, the "nickname" of the library that is used in the tables below (in quotes), and the approximate number of clones in the library.

Table 4 Description of cDNA Libraries

Library (lib #)	Description	Number of Clones in this Clustering
1	Km12 L4	

Library (lib #)	Description	Number of Clones in this Clustering
	Human Colon Cell Line, High Metastatic Potential (derived from Km12C) "High Colon"	307133
2	Km12C Human Colon Cell Line, Low Metastatic Potential "Low Colon"	284755
3	MDA-MB-231 Human Breast Cancer Cell Line, High Metastatic Potential; micro-metastases in lung "High Breast"	326937
4	MCF7 Human Breast Cancer Cell, Non Metastatic "Low Breast"	318979
8	MV-522 Human Lung Cancer Cell Line, High Metastatic Potential "High Lung"	223620
9	UCP-3 Human Lung Cancer Cell Line, Low Metastatic Potential "Low Lung"	312503
12	Human microvascular endothelial cells (HMEC) – Untreated PCR (OligodT) cDNA library	41938
13	Human microvascular endothelial cells (HMEC) – **Basic fibroblast growth factor** (bFGF) treated PCR (OligodT) cDNA library	42100
14	Human microvascular endothelial cells (HMEC) – Vascular endothelial growth factor (VEGF) treated PCR (OligodT) cDNA library	42825
15	Normal Colon – UC#2 Patient PCR (OligodT) cDNA library "Normal Colon Tumor Tissue"	34285
16	Colon Tumor – UC#2 Patient PCR (OligodT) cDNA library "Normal Colon Tumor Tissue"	35625
17	Liver Metastasis from Colon Tumor of UC#2 Patient PCR (OligodT) cDNA library "High Colon Metastasis Tissue"	36984
18	Normal Colon – UC#3 Patient PCR (OligodT) cDNA library "Normal Colon Tumor Tissue"	36216
19	Colon Tumor – UC#3 Patient PCR (OligodT) cDNA library "High Colon Tumor Tissue"	41388
20	Liver Metastasis from Colon Tumor of UC#3 Patient PCR (OligodT) cDNA library "High Colon Metastasis Tissue"	30956

The KM12L4 and KM12C cell lines are described in Example 1 above. The MDA-MB-231 cell line was originally isolated from pleural effusions (Cailleau, *J. Natl. Cancer. Inst.* (1974) *53*:661), is of high metastatic potential, and forms poorly differentiated

DOCID: <WO___9938972A2_I_>

adenocarcinoma grade II in nude mice consistent with breast carcinoma. The MCF7 cell line was derived from a pleural effusion of a breast adenocarcinoma and is non-metastatic. The MV-522 cell line is derived from a human lung carcinoma and is of high metastatic potential. The UCP-3 cell line is a low metastatic human lung carcinoma cell line; the MV-522 is a high metastatic variant of UCP-3. These cell lines are well-recognized in the art as models for the study of human breast and lung cancer (see, *e.g.*, Chandrasekaran *et al.*, *Cancer Res.* (1979) *39*:870 (MDA-MB-231 and MCF-7); Gastpar *et al.*, *J Med Chem* (1998) *41*:4965 (MDA-MB-231 and MCF-7); Ranson *et al.*, *Br J Cancer* (1998) *77*:1586 (MDA-MB-231 and MCF-7); Kuang *et al.*, *Nucleic Acids Res* (1998) *26*:1116 (MDA-MB-231 and MCF-7); Varki *et al.*, *Int J Cancer* (1987) *40*:46 (UCP-3); Varki *et al.*, *Tumour Biol.* (1990) *11*:327; (MV-522 and UCP-3); Varki *et al.*, *Anticancer Res.* (1990) *10*:637; (MV-522); Kelner *et al.*, *Anticancer Res* (1995) *15*:867 (MV-522); and Zhang *et al.*, *Anticancer Drugs* (1997) *8*:696 (MV522)). The samples of libraries 15-20 are derived from two different patients (UC#2, and UC#3). The bFGF-treated HMEC were prepared by incubation with bFGF at 10ng/ml for 2 hrs; the VEGF-treated HMEC were prepared by incubation with 20ng/ml BEGF for 2 hrs. Following incubation with the respective growth factor, the cells were washed and lysis buffer added for RNA preparation.

Each of the libraries is composed of a collection of cDNA clones that in turn are representative of the mRNAs expressed in the indicated mRNA source. In order to facilitate the analysis of the millions of sequences in each library, the sequences were assigned to clusters. The concept of "cluster of clones" is derived from a sorting/grouping of cDNA clones based on their hybridization pattern to a panel of roughly 300 7bp oligonucleotide probes (see Drmanac *et al.*, *Genomics* (1996) *37*(1):29). Random cDNA clones from a tissue library are hybridized at moderate stringency to 300 7bp oligonucleotides. Each oligonucleotide has some measure of specific hybridization to that specific clone. The combination of 300 of these measures of hybridization for 300 probes equals the "hybridization signature" for a specific clone. Clones with similar sequence will have similar hybridization signatures. By developing a sorting/grouping algorithm to analyze these signatures, groups of clones in a library can be identified and brought together computationally. These groups of clones are termed "clusters". Depending on the stringency of the selection in the algorithm (similar to the stringency of hybridization in a

classic library cDNA screening protocol), the "purity" of each cluster can be controlled. For example, artifacts of clustering may occur in computational clustering just as artifacts can occur in "wet-lab" screening of a cDNA library with 400 bp cDNA fragments, at even the highest stringency. The stringency used in the implementation of cluster herein provides groups of clones that are in general from the same cDNA or closely related cDNAs. Closely related clones can be a result of different length clones of the same cDNA, closely related clones from highly related gene families, or splice variants of the same cDNA.

Differential expression for a selected cluster was assessed by first determining the number of cDNA clones corresponding to the selected cluster in the first library (Clones in 1^{st}), and the determining the number of cDNA clones corresponding to the selected cluster in the second library (Clones in 2^{nd}). Differential expression of the selected cluster in the first library relative to the second library is expressed as a "ratio" of percent expression between the two libraries. In general, the "ratio" is calculated by: 1) calculating the percent expression of the selected cluster in the first library by dividing the number of clones corresponding to a selected cluster in the first library by the total number of clones analyzed from the first library; 2) calculating the percent expression of the selected cluster in the second library by dividing the number of clones corresponding to a selected cluster in a second library by the total number of clones analyzed from the second library; 3) dividing the calculated percent expression from the first library by the calculated percent expression from the second library. If the "number of clones" corresponding to a selected cluster in a library is zero, the value is set at 1 to aid in calculation. The formula used in calculating the ratio takes into account the "depth" of each of the libraries being compared, *i.e.*, the total number of clones analyzed in each library.

In general, a polynucleotide is said to be significantly differentially expressed between two samples when the ratio value is greater than at least about 2, preferably greater than at least about 3, more preferably greater than at least about 5 , where the ratio value is calculated using the method described above. The significance of differential expression is determined using a z score test (Zar, Biostatistical Analysis, Prentice Hall, Inc., USA, "Differences between Proportions," pp 296-298 (1974).

Example 5: Polynucleotides Differentially Expressed in High Metastatic Potential Breast Cancer Cells Versus Low Metastatic Breast Cancer Cells

A number of polynucleotide sequences have been identified that are differentially expressed between cells derived from high metastatic potential breast cancer tissue and low metastatic breast cancer cells. Expression of these sequences in breast cancer can be valuable in determining diagnostic, prognostic and/or treatment information. For example, sequences that are highly expressed in the high metastatic potential cells can be indicative of increased expression of genes or regulatory sequences involved in the metastatic process. A patient sample displaying an increased level of one or more of these polynucleotides may thus warrant more aggressive treatment. In another example, sequences that display higher expression in the low metastatic potential cells can be associated with genes or regulatory sequences that inhibit metastasis, and thus the expression of these polynucleotides in a sample may warrant a more positive prognosis than the gross pathology would suggest.

The differential expression of these polynucleotides can be used as a diagnostic marker, a prognostic marker, for risk assessment, patient treatment and the like. These polynucleotide sequences can also be used in combination with other known molecular and/or biochemical markers.

The following tables summarize polynucleotides that are differentially expressed between high metastatic potential breast cancer cells and low metastatic potential breast cancer cells.

Table 5. Differentially expressed polynucleotides: Higher expression in high metastatic potential breast cancer (lib3) relative to low metastatic breast cancer cells (lib4)

SEQ ID NOS:	Sequence Name	Cluster ID	Lib3 clones	Lib4 clones	lib3/lib4	Zscore
45	RTA00000197AR.f.12.1	3513	17	5	3.317240	2.287632
146	RTA00000185AF.a.19.2	5749	9	0	8.780930	2.629923
154	RTA00000196F.e.7.1	1039	10	2	4.878294	1.978215
159	RTA00000182AF.l.12.1	1027	41	17	2.353059	2.926571
165	RTA00000192AF.g.23.1	6455	6	0	5.853953	2.011224
174	RTA00000181AF.e.22.3	3442	17	4	4.146550	2.562391
183	RTA00000198AF.c.17.1	6923	6	0	5.853953	2.011224
364	RTA00000187AF.g.13.1	2991	10	1	9.756589	2.371428
366	RTA00000192AF.o.19.1	3549	10	1	9.756589	2.371428
387	RTA00000191AF.j.14.1	1002	42	20	2.048883	2.570309
496	RTA00000190AF.p.3.1	2378	34	0	33.17240	5.588184
510	RTA00000178AF.n.23.1	3298	12	1	11.70790	2.729313
512	RTA00000191AF.c.3.1	3549	10	1	9.756589	2.371428
529	RTA00000178AF.b.13.1	3114	9	1	8.780930	2.174815
560	RTA00000184AF.i.23.3	1577	25	3	8.130490	3.903813
606	RTA00000179AR.e.01.4	2493	33	9	3.577416	3.469507

SEQ ID NOS:	Sequence Name	Cluster ID	Lib3 clones	Lib4 clones	lib3/lib4	Zscore
644	RTA00000197F.i.12.1	3605	14	1	13.65922	3.050936
646	RTA00000186AF.d.24.1	3114	9	1	8.780930	2.174815
754	RTA00000187AF.l.11.1	4482	14	3	4.553074	2.374769
875	RTA00000401F.m.02.1	1573	34	7	4.738914	3.982056
902	RTA00000422F.c.02.1	2902	18	5	3.512372	2.443314
921	RTA00000418F.m.19.1	8890	6	0	5.853953	2.011224
942	RTA00000351R.g.11.1	3077	17	4	4.146550	2.562391
1095	RTA00000408F.l.13.1	4423	12	1	11.70790	2.729313
1104	RTA00000404F.m.10.2	779	60	22	2.660887	3.974953
1131	RTA00000400F.k.22.1	2512	7	0	6.829612	2.235371
1170	RTA00000340R.f.05.1	3202	18	3	5.853953	2.998867
1184	RTA00000422F.c.17.1	1360	26	11	2.306102	2.226876
1205	RTA00000118A.a.23.1	3500	12	3	3.902635	2.018050
1354	RTA00000401F.k.14.1	211	121	43	2.745458	5.856098
2124	RTA00000191AF.j.14.1	1002	42	20	2.048883	2.570309
1535	RTA00000405F.l.11.1	2055	29	8	3.536763	3.213373
1751	RTA00000423F.j.03.1	5391	6	0	5.853953	2.011224
1764	RTA00000399F.o.24.1	2272	17	1	16.58620	3.483575
1777	RTA00000401F.j.15.1	3061	14	0	13.65922	3.428594
1795	RTA00000348R.o.12.1	2263	6	0	5.853953	2.011224
1869	RTA00000340F.f.22.1	1720	57	8	6.951569	5.855075
1882	RTA00000401F.g.22.1	1147	28	12	2.276537	2.294031
1890	RTA00000346F.o.16.1	176	170	44	3.769591	8.366611
1915	RTA00000400F.g.02.1	1508	21	5	4.097767	2.879196
2040	RTA00000527F.j.02.2	4896	11	0	10.73224	2.974502
2059	RTA00000528F.i.22.1	2478	17	5	3.317240	2.287632
2223	RTA00000528F.j.11.1	1070	26	6	4.227855	3.289393
2245	RTA00000527F.k.09.1	213	17	4	4.146550	2.562391
2300	RTA00000528F.b.03.1	2078	11	2	5.366124	2.174565
2325	RTA00000525F.d.13.1	349	77	1	75.12573	8.384408
2462	RTA00000528F.g.22.2	920	76	32	2.317189	4.010278
2488	RTA00000528F.h.02.2	1701	18	4	4.390465	2.714073
2492	RTA00000528F.c.11.1	1701	18	4	4.390465	2.714073

Table 6: Differentially expressed polynucleotides: Higher expression in low metastatic breast cancer cells (lib4) relative to high metastatic potential breast cancer (lib3)

SEQ ID NOS:	Sequence Name	Cluster ID	Lib4 Clones	Lib 3 Clones	lib4/lib3	Zscore
15	RTA00000177AR.n.8.1	4188	4	13	3.33108	1.99126
36	RTA00000181AF.p.4.3	40392	1	8	8.19958	2.03713
44	RTA00000199F.f.08.2	12445	0	11	11.2744	3.05623
89	RTA00000177AF.n.8.3	4188	4	13	3.33108	1.99126
172	RTA00000186AF.p.09.2	6879	3	43	14.6909	5.83444
203	RTA00000201F.d.09.1	1827	37	157	4.34910	8.71727
261	RTA00000192AF.a.24.1	13183	0	7	7.17463	2.30057
419	RTA00000182AF.j.20.1	4769	2	20	10.2494	3.68254
420	RTA00000181AF.c.11.1	4769	2	20	10.2494	3.68254
503	RTA00000197AF.k.9.1	3138	1	10	10.2494	2.45316
552	RTA00000193AF.b.24.1	35	386	1967	5.22298	33.2328
564	RTA00000200AF.g.18.1	1600	0	23	23.5738	4.64683

XOCID: <WO__9938972A2_I_>

SEQ ID NOS:	Sequence Name	Cluster ID	Lib4 Clones	Lib 3 Clones	lib4/lib3	Zscore
570	RTA00000183AF.a.19.2	3788	0	6	6.14969	2.07158
590	RTA00000190AF.d.2.1	2444	26	55	2.16815	3.22244
693	RTA00000198F.m.12.1	4	987	2807	2.91492	30.3819
707	RTA00000179AF.p.15.1	5622	2	13	6.66216	2.62993
711	RTA00000198F.i.2.1	8076	0	9	9.22453	2.70385
726	RTA00000200R.f.10.1	4	987	2807	2.91492	30.3819
746	RTA00000178AF.i.01.2	4	987	2807	2.91492	30.3819
756	RTA00000404F.a.02.1	9738	1	13	13.3243	2.98623
990	RTA00000126A.o.23.1	6268	3	18	6.14969	3.11179
1122	RTA00000401F.o.06.1	2679	4	23	5.89345	3.52846
1142	RTA00000411F.a.15.1	73812	0	12	12.2993	3.21838
1286	RTA00000345F.n.12.1	7337	3	16	5.46639	2.80694
1289	RTA00000126A.g.7.1	1902	13	48	3.78442	4.45002
1435	RTA00000345F.e.11.1	4392	1	8	8.19958	2.03713
1860	RTA00000340F.p.18.1	287	6	173	29.5526	12.5749
1933	RTA00000400F.f.11.1	4088	0	82	84.0457	9.05778
1934	RTA00000341F.o.12.1	2883	9	21	2.39154	2.07600
1979	RTA00000122A.h.24.1	48	412	1020	2.53749	16.5262
1980	RTA00000346F.j.13.1	5337	5	17	3.48482	2.40321
2007	RTA00000400F.g.08.1	1275	15	32	2.18655	2.41857
2023	RTA00000523F.d.19.1	26489	1	8	8.19958	2.03713
2409	RTA00000526F.d.17.1	2757	4	16	4.09979	2.51500
1220	RTA00000528F.d.04.1	2395	12	37	3.16025	3.51521

Example 6: Polynucleotides Differentially Expressed in High Metastatic Potential Lung Cancer Cells Versus Low Metastatic Lung Cancer Cells

A number of polynucleotide sequences have been identified that are differentially expressed between cells derived from high metastatic potential lung cancer tissue and low metastatic lung cancer cells. Expression of these sequences in lung cancer tissue can be valuable in determining diagnostic, prognostic and/or treatment information. For example, sequences that are highly expressed in the high metastatic potential cells are associated can be indicative of increased expression of genes or regulatory sequences involved in the metastatic process. A patient sample displaying an increased level of one or more of these polynucleotides may thus warrant more aggressive treatment. In another example, sequences that display higher expression in the low metastatic potential cells can be associated with genes or regulatory sequences that inhibit metastasis, and thus the expression of these polynucleotides in a sample may warrant a more positive prognosis than the gross pathology would suggest.

The differential expression of these polynucleotides can be used as a diagnostic marker, a prognostic marker, for risk assessment, patient treatment and the like. These

polynucleotide sequences can also be used in combination with other known molecular and/or biochemical markers.

The following tables summarize polynucleotides that are differentially expressed between high metastatic potential lung cancer cells and low metastatic potential lung cancer cells:

Table 7 Differentially expressed polynucleotides: Higher expression in high metastatic potential lung cancer cells (lib8) relative to low metastatic lung cancer cells (lib9)

SEQ ID NO:	Sequence Name	Cluster ID	Lib8 clones	Lib9 clones	lib8/lib9	Zscore
10	RTA00000198AF.n.16.1	3721	9	0	12.5772	3.20845
54	RTA00000200F.o.22.1	983	8	1	11.1797	2.53243
65	RTA00000198AF.m.16.1	51	348	66	7.36849	17.4315
171	RTA00000198R.c.07.1	19181	6	0	8.38484	2.48169
203	RTA00000201F.d.09.1	1827	45	15	4.19242	5.09891
252	RTA00000181AF.e.18.3	8	1355	122	15.5211	39.0214
253	RTA00000181AF.e.17.3	8	1355	122	15.5211	39.0214
285	RTA00000181AR.j.14.3	5399	12	0	16.7696	3.80239
419	RTA00000182AF.j.20.1	4769	10	3	4.65824	2.29362
420	RTA00000181AF.c.11.1	4769	10	3	4.65824	2.29362
491	RTA00000196F.k.11.1	3	986	392	3.51507	22.4683
525	RTA00000198AF.c.7.1	19181	6	0	8.38484	2.48169
526	RTA00000185AF.e.20.1	5865	12	0	16.7696	3.80239
552	RTA00000193AF.b.24.1	35	868	11	110.273	34.2897
693	RTA00000198F.m.12.1	4	506	209	3.38335	15.7309
700	RTA00000183AF.i.18.2	40129	7	0	9.78231	2.74441
726	RTA00000200R.f.10.1	4	506	209	3.38335	15.7309
742	RTA00000177AF.m.1.1	14929	23	16	2.00886	2.02420
746	RTA00000178AF.i.01.2	4	506	209	3.38335	15.7309
861	RTA00000339F.f.11.1	5832	5	0	6.98736	2.18988
990	RTA00000126A.o.23.1	6268	5	0	6.98736	2.18988
1088	RTA00000399F.f.11.1	40167	8	0	11.1797	2.98512
1288	RTA00000423F.e.11.1	2566	11	2	7.68610	2.85611
1417	RTA00000339F.o.07.1	2566	11	2	7.68610	2.85611
1444	RTA00000419F.p.03.1	1937	10	3	4.65824	2.29362
1454	RTA00000340F.l.05.1	38935	7	0	9.78231	2.74441
1570	RTA00000403F.a.17.1	13686	8	0	11.1797	2.98512
1597	RTA00000401F.n.23.1	1552	8	1	11.1797	2.53243
1979	RTA00000122A.h.24.1	48	342	155	3.08345	12.2138
2024	RTA00000528F.b.23.1	1605	22	4	7.68610	4.23808
2034	RTA00000528F.m.16.1	4468	6	1	8.38484	1.97787
2126	RTA00000526F.d.01.1	4468	6	1	8.38484	1.97787

XCID: <WO__9938972A2_I_>

Table 8 Differentially expressed polynucleotides: Higher expression in low metastatic lung cancer cells (lib9) relative to high metastatic potntial lung cancer cells

SEQ ID NO:	Sequence Name	Cluster ID	Lib8 clones	Lib9 clones	lib9/lib8	Zscore
174	RTA00000181AF.e.22.3	3442	5	23	3.291654	2.368262
254	RTA00000178AF.n.2.1	17083	0	8	5.724617	2.034117
466	RTA00000177AF.p.20.1	4141	4	27	4.830145	3.070829
571	RTA00000198AF.b.14.1	801	16	46	2.057284	2.411087
574	RTA00000192AF.f.3.1	5257	5	25	3.577885	2.596857
590	RTA00000190AF.d.2.1	2444	12	37	2.206362	2.299984
922	RTA00000399F.l.14.1	3354	5	20	2.862308	1.998763
1355	RTA00000406F.m.04.1	14959	11	41	2.667151	2.865855
1422	RTA00000405F.h.07.2	4984	3	16	3.816411	2.058861
2007	RTA00000400F.g.08.1	1275	10	42	3.005423	3.147111
2038	RTA00000527F.p.06.1	1292	8	33	2.951755	2.724411
2245	RTA00000527F.k.09.1	213	137	403	2.104945	7.661033

Example 7: Polynucleotides Differentially Expressed in High Metastatic Potential Colon Cancer Cells Versus Low Metastatic Colon Cancer Cells

A number of polynucleotide sequences have been identified that are differentially expressed between cells derived from high metastatic potential colon cancer tissue and low metastatic colon cancer cells. Expression of these sequences in colon cancer tissue can be valuable in determining diagnostic, prognostic and/or treatment information. For example, sequences that are highly expressed in the high metastatic potential cells can be indicative of increased expression of genes or regulatory sequences involved in the metastatic process. A patient sample displaying an increased level of one or more of these polynucleotides may thus warrant more aggressive treatment. In another example, sequences that display higher expression in the low metastatic potential cells can be associated with genes or regulatory sequences that inhibit metastasis, and thus the expression of these polynucleotides in a sample may warrant a more positive prognosis than the gross pathology would suggest.

The differential expression of these polynucleotides can be used as a diagnostic marker, a prognostic marker, for risk assessment, patient treatment and the like. These polynucleotide sequences can also be used in combination with other known molecular and/or biochemical markers.

The following table summarizes identified polynucleotides with differential expression between high metastatic potential colon cancer cells and low metastatic potential colon cancer cells:

Table 9 Differentially expressed polynucleotides: Higher expression in high metastatic potential colon cancer (lib1) relative to low metastatic colon cancer cells (lib2)

SEQ ID NO:	Sequence Name	Cluster ID	Lib1 clones	Lib2 clones	lib1/lib2	Zscore
228	RTA00000187AR.h.15.2	6660	7	0	6.489973399	2.169320547
280	RTA00000193AF.b.18.1	7542	8	0	7.417112456	2.36964728
355	RTA00000184AR.b.24.1	5777	9	1	8.344251513	2.09555146
491	RTA00000196F.k.11.1	3	5268	2164	2.257009497	32.96556438
603	RTA00000183AR.d.11.3	6420	8	0	7.417112456	2.36964728
680	RTA00000177AF.f.10.1	6420	8	0	7.417112456	2.36964728
752	RTA00000192AF.o.7.1	5275	11	2	5.099264814	2.083995588
753	RTA00000192AF.o.17.1	5275	11	2	5.099264814	2.083995588
1241	RTA00000346F.l.13.1	7542	8	0	7.417112456	2.36964728
1264	RTA00000349R.g.10.1	5777	9	1	8.344251513	2.09555146
1401	RTA00000421F.m.14.1	3524	21	6	3.2449867	2.499690198
1442	RTA00000350R.g.10.1	9026	7	0	6.489973399	2.169320547
1514	RTA00000399F.o.06.1	13574	7	0	6.489973399	2.169320547
1851	RTA00000421F.a.06.1	2385	27	4	6.258188635	3.743586088
1915	RTA00000400F.g.02.1	1508	46	17	2.508729213	3.230059264
2024	RTA00000528F.b.23.1	1605	36	11	3.034273278	3.244010467
2066	RTA00000528F.m.12.1	5768	12	0		3.046665462

Table 10 Differentially expressed polynucleotides: Higher expression in low metastatic colon cancer cells (lib2)relative to high metastatic potential colon cancer (lib1)

SEQ ID NOS:	Sequence Name	Cluster ID	Lib1 clones	Lib2 clones	lib2/lib1	Zscore
33	RTA00000178AR.a.20.1	945	9	21	2.51670	2.21703
250	RTA00000192AF.j.21.1	2289	3	23	8.26916	3.92187
282	RTA00000193AF.c.15.1	3726	3	14	5.03340	2.58312
370	RTA00000179AF.c.15.3	2995	4	13	3.50540	2.09770
387	RTA00000191AF.j.14.1	1002	12	65	5.84234	6.26259
443	RTA00000197AR.i.17.1	3516	5	17	3.66719	2.52439
460	RTA00000179AF.c.15.1	2995	4	13	3.50540	2.09770
545	RTA00000196F.a.2.1	3575	5	14	3.02004	2.00158
560	RTA00000184AF.i.23.3	1577	12	40	3.59528	4.01991
703	RTA00000198F.l.09.1	3611	2	13	7.01081	2.73040
704	RTA00000190AF.o.12.1	3438	5	14	3.02004	2.00158
1095	RTA00000408F.l.13.1	4423	1	8	8.62869	2.11495
1104	RTA00000404F.m.10.2	779	27	54	2.15717	3.23169
1205	RTA00000118A.a.23.1	3500	3	13	4.67387	2.40298
1354	RTA00000401F.k.14.1	211	109	206	2.03843	6.08597
1387	RTA00000191AF.j.14.1	1002	12	65	5.84234	6.26259
1734	RTA00000345F.b.17.1	945	9	21	2.51670	2.21703
1742	RTA00000422F.b.22.1	2368	14	34	2.61942	3.00662
1954	RTA00000401F.j.23.1	570	59	148	2.70560	6.66631
2262	RTA00000527F.o.12.1	688	29	60	2.23155	3.53946
2325	RTA00000525F.d.13.1	349	69	138	2.15717	5.27497

Example 8: Polynucleotides Differentially Expressed in High Metastatic Potential Colon Cancer Patient Tissue Versus Normal Patient Tissue

A number of polynucleotide sequences have been identified that are differentially expressed between cells derived from high metastatic potential colon cancer tissue and normal tissue. Expression of these sequences in colon cancer tissue can be valuable in determining diagnostic, prognostic and/or treatment information. For example, sequences that are highly expressed in the high metastatic potential cells are associated can be indicative of increased expression of genes or regulatory sequences involved in the advanced disease state which involves processes such as angiogenesis, dedifferentiation, cell replication, and metastasis. A patient sample displaying an increased level of one or more of these polynucleotides may thus warrant more aggressive treatment.

The differential expression of these polynucleotides can be used as a diagnostic marker, a prognostic marker, for risk assessment, patient treatment and the like. These polynucleotide sequences can also be used in combination with other known molecular and/or biochemical markers.

The following tables summarize polynucleotides that are differentially expressed between high metastatic potential colon cancer cells and normal colon cells:

Table 11 Differentially expressed polynucleotides isolated from samples from two patients (UC#2 and UC#3) : Higher expression in high metastatic potential colon tissue (UC#2:lib17; UC#3:lib20) vs. normal colon tissue (UC#2:lib15; UC#3:lib18)

SEQ ID NO:	Sequence Name	Cluster ID	lib15 clones	lib17 clones	lib17/lib15	Zscore
65	RTA00000198AF.m.16.1	51	1	10	9.27022	2.28830
1780	RTA00000118A.j.24.1	18	4	23	5.33037	3.27028
1899	RTA00000345F.j.09.1	13	14	80	5.29727	6.34580
SEQ ID NO:	Sequence Name	Cluster ID	lib18 clones	lib20 clones	lib20/lib18	Zscore
1899	RTA00000345F.j.09.1	13	12	23	2.24234	2.16077

Table 12 Differentially expressed polynucleotides isolated from samples from two patients (UC#2 and UC#3) : Higher expression in normal colon tissue (UC#2:lib15; UC#3:lib18)vs. high metastatic potential colon tissue (UC#2:lib17; UC#3:lib20).

SEQ ID NO:	Sequence Name	Cluster ID	Lib5 Clones	L1ib7 Clones	lib15/lib17	Z Score: >2.5899%; >1.96
491	RTA00000196F.k.11.1	3	242	26	10.04	13.78900072
SEQ ID	Sequence Name	Cluster	Lib18	Lib20	lib18/lib20	Zscore

NO:		ID	clones	clones		
491	RTA00000196F.k.11.1	3	409	46	7.59993	15.3998

Example 9: Polynucleotides Differentially Expressed in High Colon Tumor Potential Patient Tissue Versus Metastasized Colon Cancer Patient Tissue

A number of polynucleotide sequences have been identified that are differentially expressed between cells derived from high tumor potential colon cancer tissue and cells derived from high metastatic potential colon cancer cells. Expression of these sequences in colon cancer tissue can be valuable in determining diagnostic, prognostic and/or treatment information associated with the transformation of precancerous tissue to malignant tissue. This information can be useful in the prevention of achieving the advanced malignant state in these tissues, and can be important in risk assessment for a patient.

The following table summarizes identified polynucleotides with differential expression between high tumor potential colon cancer tissue and cells derived from high metastatic potential colon cancer cells:

Table 13 Differentially expressed polynucleotides: High tumor potential colon tissue vs. metastatic colon tissue

SEQ ID NO:	Sequence Name	Cluster ID	L19 clones	L20 clones	lib19/lib20	Zscore
252	RTA00000181AF.e.18.3	8	14	1	10.4712	2.56699
253	RTA00000181AF.e.17.3	8	14	1	10.4712	2.56699
491	RTA00000196F.k.11.1	3	328	46	5.33318	11.8962
581	RTA00000191AF.p.3.2	17	24	2	8.97535	3.41950
693	RTA00000198F.m.12.1	4	26	8	2.43082	2.09705
726	RTA00000200R.f.10.1	4	26	8	2.43082	2.09705
746	RTA00000178AF.i.01.2	4	26	8	2.43082	2.09705
1780	RTA00000118A.j.24.1	18	80	13	4.60274	5.51440
1899	RTA00000345F.j.09.1	13	148	23	4.81287	7.68618

Example 10: Polynucleotides Differentially Expressed in High Tumor Potential Colon Cancer Patient Tissue Versus Normal Patient Tissue

A number of polynucleotide sequences have been identified that are differentially expressed between cells derived from high tumor potential colon cancer tissue and normal tissue. Expression of these sequences in colon cancer tissue can be valuable in determining diagnostic, prognostic and/or treatment information associated with the prevention of achieving the malignant state in these tissues, and can be important in risk assessment for a

DOCID: <WO___9938972A2_I_>

patient. For example, sequences that are highly expressed in the potential colon cancer cells are associated with or can be indicative of increased expression of genes or regulatory sequences involved in early tumor progression. A patient sample displaying an increased level of one or more of these polynucleotides may thus warrant closer attention or more frequent screening procedures to catch the malignant state as early as possible.

The following tables summarize polynucleotides that are differentially expressed between high metastatic potential colon cancer cells and normal colon cells:

Table 14 Differentially expressed polynucleotides detected in samples from two patients (UC#2 and UC#3): Higher expression in tumor potential colon tissue (UC#2:lib16; UC#3:lib19)vs. normal colon tissue (UC#2:lib15; UC#3:lib18)

SEQ ID NO:	Sequence Name	Cluster ID	Lib15 clones	Lib16 clones	lib16/lib15	Zscore
1899	RTA00000345F.j.09.1	13	14	50	3.43709	4.22436
SEQ ID NO:	Sequence Name	Cluster ID	Lib18 clones	Lib19 clones	lib19/lib18	Zscore
65	RTA00000198AF.m.16.1	51	0	14	12.2505	3.23250
252	RTA00000181AF.e.18.3	8	1	14	12.2505	2.84687
253	RTA00000181AF.e.17.3	8	1	14	12.2505	2.84687
581	RTA00000191AF.p.3.2	17	4	24	5.25021	3.24580
693	RTA00000198F.m.12.1	4	6	26	3.79182	2.98901
716	RTA00000200F.p.05.1	3984	0	7	6.12525	2.09621
726	RTA00000200R.f.10.1	4	6	26	3.79182	2.98901
746	RTA00000178AF.i.01.2	4	6	26	3.79182	2.98901
1780	RTA00000118A.j.24.1	18	10	80	7.00028	6.65963
1899	RTA00000345F.j.09.1	13	12	148	10.7921	9.86174

Table 15 Differentially expressed polynucleotides: Higher expression in normal colon tissue (UC#2:lib15) vs. tumor potential colon tissue (UC#2:lib16)

SEQ ID NO:	Sequence Name	Cluster ID	Lib15 clones	Lib16 clones	lib15/lib16	Zscore
491	RTA00000196F.k.11.1	3	242	39	6.44765	12.3988

Example 11: Polynucleotides Differentially Expressed in Growth Factor-Stimulated Human Microvascular Endothelial Cells (HMEC) Relative to Untreated HMEC

A number of polynucleotide sequences have been identified that are differentially expressed between human microvascular endothelial cells (HMEC) that have been treated with growth factors relative to untreated HMEC.

Sequences that are differentially expressed between growth factor-treated HMEC and untreated HMEC can represent sequences encoding gene products involved in angiogenesis, metastasis (cell migration), and other development and oncogenic processes. For example, sequences that are more highly expressed in HMEC treated with growth factors (such as bFGF or VEGF) relative to untreated HMEC can serve as markers of

cancer cells of higher metastatic potential. Detection of expression of these sequences in colon cancer tissue can be valuable in determining diagnostic, prognostic and/or treatment information associated with the prevention of achieving the malignant state in these tissues, and can be important in risk assessment for a patient. A patient sample displaying an increased level of one or more of these polynucleotides may thus warrant closer attention or more frequent screening procedures to catch the malignant state as early as possible.

The following table summarizes identified polynucleotides with differential expression between growth factor-treated and untreated HMEC.

Table 16 Differentially expressed polynucleotides: Higher expression in bFGF treated HMEC (lib13) vs. untreated HMEC (lib12)

SEQ ID NO:	Sequence Name	Cluster ID	Lib12 clones	Lib13 clones	lib13/lib12	Zscore
648	RTA00000199F.i.9.1	7	25	52	2.07199	2.94741

Table 17 Differentially expressed polynucleotides: Higher expression in VEGF treated HMEC (lib14) vs. untreated HMEC (lib12)

SEQ ID NO:	Sequence Name	Cluster ID	Lib12 clones	Lib14 clones	lib14/lib12	Zscore
648	RTA00000199F.i.9.1	7	25	67	2.62449	4.17666
1899	RTA00000345F.j.09.1	13	22	49	2.18114	2.99887

Example 12: Polynucleotides Differentially Expressed Across Multiple Libraries

A number of polynucleotide sequences have been identified that are differentially expressed between cancerous cells and normal cells across all three tissue types tested (*i.e.*, breast, colon, and lung). Expression of these sequences in a tissue or any origin can be valuable in determining diagnostic, prognostic and/or treatment information associated with the prevention of achieving the malignant state in these tissues, and can be important in risk assessment for a patient. These polynucleotides can also serve as non-tissue specific markers of, for example, risk of metastasis of a tumor. The following table summarizes identified polynucleotides that were differentially expressed but without tissue type-specificity in the breast, colon, and lung libraries tested.

Table 18 Polynucleotides Differentially Expressed Across Multiple Library Comparisons

SEQ ID NO.	Cluster	Clones in 1st Lib	Clones in 2nd Lib	Ratio	Cell or Tissue Sample and Cancer State Compared (Z Score)

DOCID: <WO___9938972A2_I_>

SEQ ID NO.	Cluster	Clones in 1st Lib	Clones in 2nd Lib	Ratio	Cell or Tissue Sample and Cancer State Compared (Z Score)
2024	1605	lib1	lib2	lib1/lib2	colon: high met > low met
		36	11	3.0342732	(3.2440104)
		lib8	lib9	lib8/lib9	lung: high met > low met
		22	4	7.6861036	(4.2380835)
65	51	lib8	lib9	lib8/lib9	lung: high met > low met
		348	66	7.3684960	(17.431560)
		lib18	lib19	lib19/lib18	pt #3 colon: tumor > normal
		0	14	12.250507	(3.2325073)
		lib15	lib17	lib17/lib15	pt #2 colon: met > normal
		1	10	9.2702249	(2.2883061)
174	3442	lib8	lib9	lib9/lib8	lung: low met > high met
		5	23	3.2916548	(2.3682625)
		lib3	lib4	lib3/lib4	breast: high met > low met
		17	4	4.1465504	(2.5623912)
203	1827	lib8	lib9	lib8/lib9	lung: high met > low met
		45	15	4.1924201	(5.0989192)
		lib3	lib4	lib4/lib3	breast: low met > high met
		37	157	4.3491051	(8.7172773)
2245	213	lib8	lib9	lib9/lib8	lung: low met > high met
		137	403	2.1049458	(7.6610331)
		lib3	lib4	lib3/lib4	breast: high met > low met
		17	4	4.1465504	(2.5623912)
990	6268	lib8	lib9	lib8/lib9	lung: high met > low met
		5	0	6.9873669	(2.1898837)
		lib3	lib4	lib4/lib3	breast: low met > high met
		3	18	6.1496901	(3.1117967)
252	8	lib8	lib9	lib8/lib9	lung: high met > low met
		1355	122	15.521118	(39.021411)
		lib19	lib20	lib19/lib20	pt. #3 colon: tumor > met
		14	1	10.471247	(2.5669948)
		lib18	lib19	lib19/lib18	pt #3 colon: tumor > normal
		1	14	12.250507	(2.8468716)
253	8	lib8	lib9	lib8/lib9	lung: high met > low met
		1355	122	15.521118	(39.021411)
		lib19	lib20	lib19/lib20	pt. #3 colon: tumor > met
		14	1	10.471247	(2.5669948)
		lib18	lib19	lib19/lib18	pt #3 colon: tumor > normal
		1	14	12.250507	(2.8468716)
2325	349	lib3	lib4	lib3/lib4	breast: high met > low met
		77	1	75.125736	(8.3844087)
		lib1	lib2	lib2/lib1	colon: low met > high met
		69	138	2.1571737	(5.2749799)

SEQ ID NO.	Cluster	Clones in 1st Lib	Clones in 2nd Lib	Ratio	Cell or Tissue Sample and Cancer State Compared (Z Score)
1095	4423	lib3	lib4	lib3/lib4	breast: high met > low met
		12	1	11.707907	(2.7293134)
		lib1	lib2	lib2/lib1	colon: low met > high met
		1	8	8.6286948	(2.1149516)
1124	779	lib3	lib4	lib3/lib4	breast: high met > low met
		60	22	2.6608879	(3.9749537)
		lib1	lib2	lib2/lib1	colon: low met > high met
		27	54	2.1571737	(3.2316908)
387	1002	lib3	lib4	lib3/lib4	breast: high met > low met
		42	20	2.0488837	(2.5703094)
		lib1	lib2	lib2/lib1	colon: low met > high met
		12	65	5.8423454	(6.2625969)
419	4769	lib8	lib9	lib8/lib9	lung: high met > low met
		10	3	4.6582446	(2.2936274)
		lib3	lib4	lib4/lib3	breast: low met > high met
		2	20	10.249483	(3.6825426)
420	4769	lib8	lib9	lib8/lib9	lung: high met > low met
		10	3	4.6582446	(2.2936274)
		lib3	lib4	lib4/lib3	breast: low met > high met
		2	20	10.249483	(3.6825426)
1205	3500	lib3	lib4	lib3/lib4	breast: high met > low met
		12	3	3.9026356	(2.0180506)
		lib1	lib2	lib2/lib1	colon: low met > high met
		3	13	4.6738763	(2.4029818)
491	3	lib1	lib2	lib1/lib2	colon: high met > low met
		5268	2164	2.2570094	(32.965564)
		lib8	lib9	lib8/lib9	lung: high met > low met
		986	392	3.5150733	(22.468331)
		lib19	lib20	lib19/lib20	pt #3 colon: tumor > met
		328	46	5.3331820	(11.896271)
		lib18	lib20	lib18/lib20	pt #3 colon: normal > met
		409	46	7.5999342	(15.399861)
		lib15	lib17	lib15/lib17	pt#2 colon: normal > met
		242	26	10.04	(13.789000)
		lib15	lib16	lib15/lib16	pt#2 colon: normal > tumor
		242	39	6.44765	12.39883
552	35	lib8	lib9	lib8/lib9	lung: high met > low met
		868	11	110.27335	(34.289704)
		lib3	lib4	lib4/lib3	breast: low met > high met
		386	1967	5.2229880	(33.232871)
560	1577	lib3	lib4	lib3/lib4	breast: high met > low met
		25	3	8.1304909	(3.9038139)

OCID: <WO___9938972A2_I_>

SEQ ID NO.	Cluster	Clones in 1st Lib	Clones in 2nd Lib	Ratio	Cell or Tissue Sample and Cancer State Compared (Z Score)
		lib1	lib2	lib2/lib1	colon: low met > high met
		12	40	3.5952895	(4.0199130)
581	17	lib19	lib20	lib19/lib20	pt #3 colon: tumor > met
		24	2	8.9753551	(3.4195074)
		lib18	lib19	lib19/lib18	pt #3 colon: tumor > normal
		4	24	5.2502174	(3.2458055)
590	2444	lib3	lib4	lib4/lib3	breast: low met > high met
		26	55	2.1681599	(3.2224421)
		lib8	lib9	lib9/lib8	lung: low met > high met
		12	37	2.2063628	(2.2999846)
1354	211	lib3	lib4	lib3/lib4	breast: high met > low met
		121	43	2.7454588	(5.8560985)
		lib1	lib2	lib2/lib1	colon: low met > high met
		109	206	2.0384302	(6.0859794)
1387	1002	lib3	lib4	lib3/lib4	breast: high met > low met
		42	20	2.0488837	(2.5703094)
		lib1	lib2	lib2/lib1	colon: low met > high met
		12	65	5.8423454	(6.2625969)
648	7	lib12	lib14	lib14/lib12	HMEC: VEGF > untreated
		25	67	2.6244913	(4.1766696)
		lib12	lib13	lib13/lib12	HMEC: bFGF > untreated
		25	52	2.0719962	(2.9474155)
693	4	lib8	lib9	lib8/lib9	lung: high met > low met
		506	209	3.3833566	(15.730912)
		lib3	lib4	lib4/lib3	breast: low met > high met
		987	2807	2.9149240	(30.381945)
		lib19	lib20	lib19/lib20	pt#3 colon: tumor > met
		26	8	2.4308253	(2.0970580)
		lib18	lib19	lib19/lib18	pt#3 colon: tumor > normal
		6	26	3.7918237	(2.9890107)
726	4	lib8	lib9	lib8/lib9	lung: high met > low met
		506	209	3.3833566	(15.730912)
		lib3	lib4	lib4/lib3	breast: low met > high met
		987	2807	2.9149240	(30.381945)
		lib19	lib20	lib19/lib20	pt#3 colon: tumor > met
		26	8	2.4308253	(2.0970580)
		lib18	lib19	lib19/lib18	pt#3 colon: tumor > normal
		6	26	3.7918237	(2.9890107)
746	4	lib8	lib9	lib8/lib9	lung: high met > low met
		506	209	3.3833566	(15.730912)
		lib3	lib4	lib4/lib3	breast: low met > high met
		987	2807	2.9149240	(30.381945)

SEQ ID NO.	Cluster	Clones in 1st Lib	Clones in 2nd Lib	Ratio	Cell or Tissue Sample and Cancer State Compared (Z Score)
		lib19	lib20	lib19/lib20	pt#3 colon: tumor > met
		26	8	2.4308253	(2.0970580)
		lib18	lib19	lib19/lib18	pt#3 colon: tumor > normal
		6	26	3.7918237	(2.9890107)
1780	18	lib19	lib20	lib19/lib20	pt#3 colon: tumor > met
		80	13	4.6027462	(5.5144093)
		lib18	lib19	lib19/lib18	pt#3 colon: tumor > normal
		10	80	7.0002899	(6.6596394)
		lib15	lib17	lib17/lib15	pt#3 colon: met > normal
		4	23	5.3303793	(3.2702852)
1899	13	lib19	lib20	lib19/lib20	pt#3 colon: tumor > met
		148	23	4.8128716	(7.6861840)
		lib18	lib20	lib20/lib18	pt#3 colon: met > normal
		12	23	2.2423439	(2.1607719)
		lib18	lib19	lib19/lib18	pt#3 colon: tumor > normal
		12	148	10.792113	(9.8617485)
		lib15	lib17	lib17/lib15	pt#2 colon: met > normal
		14	80	5.2972714	(6.3458044)
		lib15	lib16	lib16/lib15	pt#2 colon: tumor > normal
		14	50	3.4370927	(4.2243697)
		lib12	lib14	lib14/lib12	HMEC: VEGF > untreated
		22	49	2.1811410	(2.9988774)
1915	1508	lib1	lib2	lib1/lib2	colon: high met > low met
		46	17	2.5087292	(3.2300592)
		lib3	lib4	lib3/lib4	breast: high met > low met
		21	5	4.0977674	(2.8791960)
1979	48	lib8	lib9	lib8/lib9	lung: high met > low met
		342	155	3.0834574	(12.213852)
		lib3	lib4	lib4/lib3	breast: low met > high met
		412	1020	2.5374934	(16.526285)
2007	1275	lib3	lib4	lib4/lib3	breast: low met > high met
		15	32	2.1865564	(2.4185764)
		lib8	lib9	lib9/lib8	lung: low met > high met
		10	42	3.0054239	3.1471113

high met = high metastatic potential; low met = low metastatic potential;
met = metastasized; tumor = non-metastasized tumor;
pt = patient; #2 = UC#2; #3 = UC#3;
HMEC = human microvascular endothelial cell;
bFGF = bFGF treated; VEGF = VEGF treated

DCID: <WO___9938972A2_I_>

Example 12: Polynucleotides Exhibiting Colon-Specific Expression

The cDNA libraries described herein were also analyzed to identify those polynucleotides that were specifically expressed in colon cells or tissue, *i.e.*, the polynucleotides were identified in libraries prepared from colon cell lines or tissue, but not in libraries of breast or lung origin. The polynucleotides that were expressed in a colon cell line and/or in colon tissue, but were present in the breast or lung cDNA libraries described herein, are shown in Table 19 (inserted before claims).

No clones corresponding to the colon-specific polynucleotides in the table above were present in any of Libraries 3, 4, 8, 9, 12, 13, 14, or 15. The polynucleotide provided above can be used as markers of cells of colon origin, and find particular use in reference arrays, as described above.

Example 13: Identification of Contiguous Sequences Having a Polynucleotide of the Invention

The novel polynucleotides were used to screen publicly available and proprietary databases to determine if any of the polynucleotides of SEQ ID NOS:1-2502 would facilitate identification of a contiguous sequence, *e.g.*, the polynucleotides would provide sequence that would result in 5' extension of another DNA sequence, resulting in production of a longer contiguous sequence composed of the provided polynucleotide and the other DNA sequence(s). Contiging was performed using the Gelmerge application (default settings) of GCG from the Univ. of Wisconsin.

Using these parameters, 146 contiged sequences were generated. These contiged sequences are provided as SEQ ID NOS:5107-5252 (see Table 1). The contiged sequences can be correlated with the sequences of SEQ ID NOS:1-2502 upon which the contiged sequences are based by, for example, identifying those sequences of SEQ ID NOS:1-2502 and the contiged sequences of SEQ ID NOS:5107-5252 that share the same clone name in Table 1.

The contiged sequences (SEQ ID NO:5107-5252) thus represent longer sequences that encompass a polynucleotide sequence of the invention. The contiged sequences were then translated in all three reading frames to determine the best alignment with individual sequences using the BLAST programs as described above for SEQ ID NOS:1-2502 and the validation sequences "SEQ ID NOS:2503-5106." Again the sequences were masked using the XBLAST program for masking low complexity as described above in Example 1

(Table 2). Several of the contiged sequences were found to encode polypeptides having characteristics of a polypeptide belonging to a known protein families (and thus represent new members of these protein families) and/or comprising a known functional domain (Table 20). Thus the invention encompasses fragments, fusions, and variants of such polynucleotides that retain biological activity associated with the protein family and/or functional domain identified herein.

Table 20 Profile hits using contiged sequences

SEQ ID NO	Biological Activity (Profile)	Start	Stop	Score	Direction	Sequence Name
5111	7tm_2	71	915	8090	for	RTA00000399F.o.01.1
5120	7tm_2	101	919	8475	rev	RTA00000341F.m.21.1
5174	7tm_2	3	963	9431	for	RTA00000192AF.h.19.1
5197	7tm_2	214	1073	8528	rev	RTA00000192AF.f.3.1
5208	ANK	546	629	4920	for	RTA00000190AF.f.5.1
5120	asp	126	1067	6620	rev	RTA00000341F.m.21.1.
5241	asp	112	1094	6553	for	RTA00000418F.i.06.1
5243	asp	347	1028	5981	for	RTA00000339F.b.02.1
5197	ATPases	113	781	5690	for	RTA00000192AF.f.3.1
5239	ATPases	1	348	15955	for	RTA00000401F.m.07.1
5241	ATPases	110	823	6782	for	RTA00000418F.i.06.1
5243	ATPases	338	874	5832	for	RTA00000339F.b.02.1
5125	protkinase	59	685	5791	for	RTA00000182AF.c.5.1
5217	protkinase	75	1035	5405	for	RTA00000181AF.p.12.3
5237	protkinase	25	546	5107	rev	RTA00000118A.n.5.1
5248	protkinase	14	422	5103	rev	RTA00000419F.k.05.1
5252	protkinase	89	755	5499	for	RTA00000404F.m.17.2
5120	Wnt_dev_sign	3	948	11036	for	RTA00000341F.m.21.1

All stop/start sequences are provided in the forward direction.

Descriptions of the profiles for the indicated protein families and functional domains are provided in Example 3 above.

Those skilled in the art will recognize, or be able to ascertain, using not more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such specific embodiments and equivalents are intended to be encompassed by the following claims.

OCID: <WO___9938972A2_I_>

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. The citation of any publication is for its disclosure prior to the filing date and should not be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it is readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

Deposit Information:

The following materials were deposited with the American Type Culture Collection: CMCC = (Chiron Master Culture Collection)

Cell Lines Deposited with ATCC

Cell Line	Deposit Date	ATCC Accession No.	CMCC Accession No.
KM12L4-A	March 19, 1998	CRL-12496	11606
Km12C	May 15, 1998	CRL-12533	11611
MDA-MB-231	May 15, 1998	CRL-12532	10583
MCF-7	October 9, 1998	CRL-12584	10377

cDNA Libraries Deposited with ATCC

cDNA Library No. Deposit Date ATCC Accession No.	cDNA Library ES21 January 22, 1999 ATCC No.	cDNA Library ES22 January 22, 1999 ATCC No.	cDNA Library ES23 January 22, 1999 ATCC No.
Clone Names	M00001575D:G05	M00001364A:E11	M00001489B:A06
	M00001460A:A03	M00001694C:H10	M00001585A:D06
	M00001655C:E04	M00003841D:E03	M00001637B:E07
	M00001676C:C11	M00004176D:B12	M00001529D:H02
	M00001679D:D05	M00001387B:E02	M00001500C:C08
	M00001546B:C05	M00004282B:A04	M00001483B:D03
	M00001453B:E10	M00001376B:F03	M00001623C:H07
		M00001445D:A06	M00003975B:F03
		M00001399C:H12	
		M00004208D:H08	

cDNA Library No.	cDNA Library ES24	cDNA Library ES25	cDNA Library ES26
Deposit Date	January 22, 1999	January 22, 1999	January 22, 1999
ATCC Accession No.	ATCC No.	ATCC No.	ATCC No.
Clone Names	M00003987D:D06	M00001675D:B08	M00001479C:F10
	M00004073A:H12	M00001589B:E12	M00003842D:F08
	M00004104B:F11	M00001607D:A11	M00003901A:C09
	M00004237D:D08	M00001636A:E07	M00003982A:B06
	M00004111D:B07	M00001530A:B12	M00003824A:A06
	M00004138B:B11	M00001495B:B08	M00003845D:C03
	M00001391C:C04	M00001487C:F01	M00003856A:B07
	M00001448D:E12	M00001644B:D06	M00004104B:A02
	M00001450A:B03	M00003751C:A04	M00004110C:E03
	M00001451B:F01		

In addition, libraries of selected clones were deposited. The details of these deposits are provided in Tables 21-24.

This deposit is provided merely as convenience to those of skill in the art, and is not an admission that a deposit is required under 35 U.S.C. §112. The sequence of the polynucleotides contained within the deposited material, as well as the amino acid sequence of the polypeptides encoded thereby, are incorporated herein by reference and are controlling in the event of any conflict with the written description of sequences herein. A license may be required to make, use, or sell the deposited material, and no such license is granted hereby.

Retrieval of Individual Clones from Deposit of Pooled Clones

Where the ATCC deposit is composed of a pool of cDNA clones, the deposit was prepared by first transfecting each of the clones into separate bacterial cells. The clones were then deposited as a pool of equal mixtures in the composite deposit. Particular clones can be obtained from the composite deposit using methods well known in the art. For example, a bacterial cell containing a particular clone can be identified by isolating single colonies, and identifying colonies containing the specific clone through standard colony hybridization techniques, using an oligonucleotide probe or probes designed to specifically hybridize to a sequence of the clone insert (*e.g.*, a probe based upon unmasked sequence of the encoded polynucleotide having the indicated SEQ ID NO). The probe should be designed to have a T_m of approximately 80°C (assuming 2°C for each A or T and 4°C for each G or C). Positive colonies can then be picked, grown in culture, and the recombinant clone isolated. Alternatively, probes designed in this manner can be used to PCR to isolate a nucleic acid molecule from the pooled clones according to methods well known in the art,

XOCID: <WO__9938972A2_I_>

e.g., by purifying the cDNA from the deposited culture pool, and using the probes in PCR reactions to produce an amplified product having the corresponding desired polynucleotide sequence.

Table 1.

SEQ ID NO:	Filing Date of Priority Appln	SEQ ID NO: in Priority Appln	Sequence Name	Clone Name	Cluster ID
1	1/28/98	1	RTA00000197AF.i.16.1	M00001490A:D11	16402
2	1/28/98	2	RTA00000188AF.n.15.1	M00003804A:H04	0
3	1/28/98	3	RTA00000197AF.e.24.1	M00001456B:F10	39250
4	1/28/98	4	RTA00000198R.f.04.1	M00001607D:F07	5023
5	1/28/98	5	RTA00000195R.c.11.1	M00003811A:E03	66087
6	1/28/98	6	RTA00000195AF.c.16.1	M00003829C:A11	23508
7	1/28/98	7	RTA00000197AR.e.12.1	M00001454B:G07	22095
8	1/28/98	8	RTA00000200AF.h.11.2	M00004146A:C08	8399
9	1/28/98	9	RTA00000177AF.g.22.1	M00001347C:G08	7031
10	1/28/98	10	RTA00000198AF.n.16.1	M00001694C:H10	3721
11	1/28/98	11	RTA00000199AF.i.17.1	M00003880C:F10	9615
12	1/28/98	12	RTA00000183AF.i.15.2	M00001529B:C04	2642
13	1/28/98	13	RTA00000190AF.i.5.1	M00003919A:A10	0
14	1/28/98	14	RTA00000196R.c.11.2	M00001352A:E12	13658
15	1/28/98	15	RTA00000177AR.n.8.1	M00001356D:F06	4188
16	1/28/98	16	RTA00000196AF.e.16.1	M00001363C:H02	39252
17	1/28/98	17	RTA00000183AR.e.14.2	M00001506B:D09	17437
18	1/28/98	18	RTA00000196AF.c.17.1	M00001352C:F06	39602
19	1/28/98	19	RTA00000185AF.a.8.1	M00001570D:A03	4868
20	1/28/98	20	RTA00000181AF.l.14.2	M00001454D:D06	2364
21	1/28/98	21	RTA00000131A.g.19.2	M00001449A:G10	36535
22	1/28/98	22	RTA00000187AR.o.10.2	M00001718D:F07	8984
23	1/28/98	23	RTA00000198R.b.08.1	M00001567C:H12	22636
24	1/28/98	24	RTA00000200AF.f.11.1	M00004111D:D11	0
25	1/28/98	25	RTA00000196AF.c.1.1	M00001349C:C05	8171
26	1/28/98	26	RTA00000200R.g.09.1	M00004131B:H09	22785
27	1/28/98	27	RTA00000192AF.i.12.1	M00004169C:C12	5319
28	1/28/98	28	RTA00000178AR.o.01.5	M00001387B:H07	0
29	1/28/98	29	RTA00000200AF.b.19.1	M00004042D:H02	22847
30	1/28/98	30	RTA00000184AR.n.07.2	M00001561C:F06	0
31	1/28/98	31	RTA00000200F.m.15.1	M00004236C:D10	22601
32	1/28/98	32	RTA00000198R.m.19.1	M00001680D:D02	40041
33	1/28/98	33	RTA00000178AR.a.20.1	M00001362C:H11	945
34	1/28/98	34	RTA00000197AF.n.8.1	M00001536D:A12	4101
35	1/28/98	35	RTA00000191AF.n.17.1	M00004091B:D11	7848
36	1/28/98	36	RTA00000181AF.p.4.3	M00001460A:A03	40392
37	1/28/98	37	RTA00000181AF.n.15.2	M00001457A:B07	86128
38	1/28/98	38	RTA00000196R.k.07.1	M00001399C:D09	22443
39	1/28/98	39	RTA00000189AR.b.19.1	M00003832B:E01	5294
40	1/28/98	40	RTA00000200AR.e.02.1	M00004090A:F09	36059
41	1/28/98	41	RTA00000184F.k.12.1	M00001557D:D09	8761
42	1/28/98	42	RTA00000184F.j.21.1	M00001557A:D02	7065
43	1/28/98	43	RTA00000179AF.c.14.3	M00001392D:H04	0
44	1/28/98	44	RTA00000199F.f.08.2	M00003841D:E03	12445
45	1/28/98	45	RTA00000197AR.f.12.1	M00001458C:E01	3513
46	1/28/98	46	RTA00000182AF.f.13.1	M00001470C:B10	8010
47	1/28/98	47	RTA00000192AF.m.12.1	M00004191D:B11	0
48	1/28/98	48	RTA00000177AR.a.23.5	M00001339D:G02	6995
49	1/28/98	49	RTA00000198R.o.05.1	M00003750A:D01	26702

DOCID: <WO___9938972A2_I_>

SEQ ID NO:	Filing Date of Priority Appln	SEQ ID NO: in Priority Appln	Sequence Name	Clone Name	Cluster ID
1038	2/24/98	283	RTA00000346F.n.06.1	M00004139C:A12	12439
1039	2/24/98	284	RTA00000412F.l.21.1	M00004029C:G10	65183
1040	2/24/98	285	RTA00000413F.i.02.1	M00004110D:A10	65857
1041	2/24/98	286	RTA00000404F.i.19.1	M00001625B:C10	38698
1042	2/24/98	287	RTA00000410F.n.09.1	M00001662C:A04	11736
1043	2/24/98	288	RTA00000403F.a.11.1	M00001448C:F10	73109
1044	2/24/98	289	RTA00000420F.n.08.1	M00005257A:H11	0
1045	2/24/98	290	RTA00000411F.k.16.1	M00003852C:B06	64759
1046	2/24/98	291	RTA00000405F.c.01.1	M00001657D:A04	19236
1047	2/24/98	292	RTA00000423F.i.18.1	M00003918A:D08	14996
1048	2/24/98	293	RTA00000403F.l.04.1	M00001571C:A04	39278
1049	2/24/98	294	RTA00000405F.l.17.1	M00003805A:F02	17225
1050	2/24/98	295	RTA00000406F.a.07.1	M00003856C:H09	26607
1051	2/24/98	296	RTA00000347F.d.06.1	M00001457C:F02	39122
1052	2/24/98	297	RTA00000419F.b.18.1	M00003808D:D08	67034
1053	2/24/98	298	RTA00000406F.h.07.1	M00003901B:H04	38003
1054	2/24/98	299	RTA00000405F.l.15.1	M00001694A:E03	19575
1055	2/24/98	300	RTA00000406F.g.17.1	M00003881B:F10	37979
1056	2/24/98	301	RTA00000401F.m.23.1	M00003914C:C02	2801
1057	2/24/98	302	RTA00000356R.f.18.1	M00004692A:H10	0
1058	2/24/98	303	RTA00000130A.h.22.1	M00001617A:D06	80933
1059	2/24/98	304	RTA00000403F.n.18.2	M00001577D:H06	8811
1060	2/24/98	305	RTA00000418F.p.06.1	M00001664A:F08	32628
1061	2/24/98	306	RTA00000404F.d.13.1	M00001595D:A04	39036
1062	2/24/98	307	RTA00000420F.l.12.2	M00005230B:H09	0
1063	2/24/98	308	RTA00000353R.d.11.1	M00004692A:H08	0
1064	2/24/98	309	RTA00000340F.n.01.1	M00001679A:G06	39081
1065	2/24/98	310	RTA00000419F.d.06.1	M00003820B:D07	65496
1066	2/24/98	311	RTA00000419F.n.09.1	M00003977C:A06	66070
1067	2/24/98	312	RTA00000399F.i.08.1	M00001575D:B10	38927
1068	2/24/98	313	RTA00000406F.g.07.1	M00003880C:E11	37925
1069	2/24/98	314	RTA00000423F.g.13.1	M00003905A:E07	38028
1070	2/24/98	315	RTA00000419F.p.12.1	M00004037A:E04	13767
1071	2/24/98	316	RTA00000414F.a.02.1	M00005178D:H04	0
1072	2/24/98	317	RTA00000195AF.b.21.1	M00001595B:A09	39055
1072	1/28/98	602	RTA00000195AF.b.21.1	M00001595B:A09	39055
1073	2/24/98	318	RTA00000403F.h.05.1	M00001482D:A04	39096
1074	2/24/98	319	RTA00000420F.b.21.1	M00004088D:B10	65057
1075	2/24/98	320	RTA00000422F.p.07.2	M00001661A:E06	39024
1076	2/24/98	321	RTA00000339F.c.21.1	M00001389C:A08	5325
1077	2/24/98	322	RTA00000339F.c.24.1	M00001364B:B06	5516
1078	2/24/98	323	RTA00000421F.n.19.1	M00001679A:D10	16409
1079	2/24/98	324	RTA00000340F.p.17.1	M00003750C:H05	0
1080	2/24/98	325	RTA00000345F.k.21.1	M00001464B:C11	40204
1081	2/24/98	326	RTA00000419F.b.15.1	M00003806D:D11	43969
1082	2/24/98	327	RTA00000405F.a.11.1	M00001655A:B11	39124
1083	2/24/98	328	RTA00000423F.k.19.2	M00003985D:E10	17615
1084	2/24/98	329	RTA00000413F.e.16.1	M00004093C:C02	63836
1085	2/24/98	330	RTA00000403F.i.04.1	M00001485B:D09	8930
1086	2/24/98	331	RTA00000404F.o.18.2	M00001651C:C05	39110

We Claim:

1. A library of polynucleotides, the library comprising the sequence information of at least one of SEQ ID NOS:1-3544, 3546-4510, 4512-4725, 4727-4748, and 4750-5252.

2. The library of claim 1, wherein the library is provided on a nucleic acid array.

3. The library of claim 1, wherein the library is provided in a computer-readable format.

4. The library of claim 1, wherein the library comprises a differentially expressed polynucleotide comprising a sequence selected from the group consisting of SEQ ID NOS:65, 174, 203, 252, 253, 387, 419, 420, 491, 552, 560, 581, 590, 648, 693, 726, 746, 990, 1095, 1124, 1205, 1354, 1387, 1780, 1899, 1915, 1979, 2007, 2024, 2245, and 2325.

5. The library of claim 1, wherein the library comprises a polynucleotide differentially expressed in a human breast cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS:15, 36, 44, 45, 89, 146, 154, 159, 165, 174, 172, 183, 203, 261, 364, 366, 387, 419, 420, 496, 503, 510, 512, 529, 552, 560, 564, 570, 590, 606, 644, 646, 693, 707, 711, 726, 746, 754, 756, 875, 902, 921, 942, 990, 1095, 1104, 1122, 1131, 1142, 1170, 1184, 1205, 1286, 1289, 1354, 1387, 1435, 1535, 1751, 1764, 1777, 1795, 1860, 1869, 1882, 1890, 1915, 1933, 1934, 1979, 1980, 2007, 2023, 2040, 2059, 2223, 2245, 2300, 2325, 2409, 2462, 2486, 2488, and 2492.

6. The library of claim 1, wherein the library comprises a polynucleotide differentially expressed in a human colon cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS:33, 65, 228, 250, 252, 253, 280, 282, 355, 370, 387, 443, 460, 491, 545, 560, 581, 603, 680, 693, 703, 704, 716, 726, 746, 752, 753, 1095, 1104, 1205, 1241, 1264, 1354, 1387, 1401, 1442, 1514, 1734, 1742, 1780, 1851, 1899, 1915, 1954, 2024, 2066, 2262, and 2325.

XOCID: <WO__9938972A2_I_>

7. The library of claim 1, wherein the library comprises a polynucleotide differentially expressed in a human lung cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS: 10, 54, 65, 171, 174, 203, 252, 253, 254, 285, 419, 420, 466, , 491, 525, 526, 552, 571, 574, 590, 693, 700, 726, 742, 746, 861, 990, 922, 1088, 1288, 1355, 1417, 1422, 1444, 1454, 1570, 1597, 1979, 2007, 2024, 2034, 2038, 2126, and 2245.

8. The library of claim 1, wherein the library comprises a polynucleotide differentially expressed in a human cancer cell, where the polynucleotide comprises a sequence selected from the group consisting of SEQ ID NOS:648 and1899.

9. An isolated polynucleotide comprising a nucleotide sequence having at least 90% sequence identity to an identifying sequence of SEQ ID NOS:1-3544, 3546-4510, 4512-4725, 4727-4748, and 4750-5252, or a degenerate variant or fragment thereof.

10. The polynucleotide of claim 9, wherein the polynucleotide comprises a sequence of one of SEQ ID NOS:2503, 2504, 2550, 2555, 2578, 2656, 2667, 2712, 2723, 2728, 2738, 2734, 2754, 2758, 2760, 2832, 2835, 2842, 2843, 2849, 2893, 2933, 2956, 2971, 2981, 3009, 3018, 3019, 3046, 3084, 3190, 3129, 3173, 3226, 3227, 3274, 3290, 3356, 3365, 3377, 3381, 3390, 3391, 3404, 3407, 3408, 3409, 3418, 3419, 3451, 3597, 3600, 3618, 3632, 3635, 3646, 3648, 3657, 3665, 3669, 3670, 3671, 3656, 3680, 3686, 3695, 3696, 3700, 3710, 3736, 3762, 3763, 3774, 3775, 3791, 3804, 3806, 3836, 3895, 3905, 3919, 3920, 3927, 3936, 3951, 3974, 3998, 4036, 4038, 4044, 4056, 4072, 4117, 4119, 4152, 4153, 4154, 4172, 4175, 4159, 4175, 4205, 4216, 4223, 4228, 4238, 4241, 4243, 4251, 4253, 4261, 4263,4278, 4288, 4322, 4330, 4343, 4359, 4363, 4364, 4365, 4373, 4375, 4384, 4385, 4406, 4409, 4431, 4434, 4441, 4442, 4444, 4455, 4469, 4473, 4477, 4482, 4489, 4495, 4496, 4498, 4525, 4535, 4536, 4540, 4560, 4616, 4562, 4586, 4605, 4629, 4653, 4654, 4658, 4659, 4660, 4661, 4664, 4665, 4668, 4684, 4682, 4688, 4689, 4710, 4718, 4733, 4724, 4733, 4746, 4755, 4760, 4710, 4777, 4785, 4792, 4794, 4801, 4807, 4821, 4822, 4847, 4850, 4854, 4856, 4866, 4885, 4900, 4901, 4905, 4914, 4925, 4929, 4931, 4943, 4944, 4959, 5111, 5020, 5041, 5046, 5059, 5083, 5090, 5094, 5102, 5125, 5174, 5197, 5208, 5217, 5237, 5239, 5241, 5243, 5248, and 5252.

BNS page 656

11. A recombinant host cell containing the polynucleotide of claim 9.

12. An isolated polypeptide encoded by the polynucleotide of claim 9.

13. An antibody that specifically binds a polypeptide of claim 12.

14. A vector comprising the polynucleotide of claim 9.

15. A polynucleotide comprising the nucleotide sequence of an insert contained in a clone deposited as ATCC accession number xx, xx, xx, xx, xx, xx, xx, xx, or xx.

16. A method of detecting differentially expressed genes correlated with a cancerous state of a mammalian cell, the method comprising the step of:

detecting at least one differentially expressed gene product in a test sample derived from a cell suspected of being cancerous, where the gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:10, 15, 33, 36, 44, 45, 54, 65, 89, 146, 154, 159, 165, 171, 172, 174, 183, 203, 228, 250, 252, 253, 254, 261, 280, 282, 285, 355, 364, 366, 370, 387, 419, 420, 443, 460, 466, 491, 496, 503, 510, 512, 525, 526, 529, 545, 552, 560, 564, 570, 571, 574, 581, 590, 603, 606, 644, 646, 648, 680, 693, 700, 703, 704, 707, 711, 716, 726, 742, 746, 752, 753, 754, 756, 861, 875, 902, 921, 922, 942, 990, 1088, 1095, 1104, 1122, 1131, 1142, 1170, 1184, 1205, 1286, 1288, 1289, 1354, 1355, 1387, 1417, 1435, 1444, 1454, 1535, 1570, 1597, 1734, 1742, 1751, 1764, 1777, 1780, 1795, 1860, 1869, 1882, 1890, 1899, 1915, 1933, 1934, 1954, 1979, 1980, 2007, 2023, 2024, 2034, 2040, 2059, 2126, 2223, 2245, 2262, 2300, 2325, 2409, 2486, 2462, 2488, 2492, 1241, 1264, 1401, 1422, 1442, 1514, 1851, 1915, 2007, 2024, 2038, 2066, and 2245;

wherein detection of the differentially expressed gene product is correlated with a cancerous state of the cell from which the test sample was derived.

17. The method of claim 16, wherein said detecting step is by hybridization of the test sample to a reference array, wherein the reference array comprises an identifying sequence of at least one of SEQ ID NOS: 65, 174, 203, 252, 253, 387, 419, 420, 491, 552,

XCID: <WO___9938972A2_I_>

560, 581, 590, 648, 693, 726, 746, 990, 1095, 1124, 1205, 1354, 1387, 1780, 1899, 1915, 1979, 2007, 2024, 2325, and 2245.

18. The method of claim 16, wherein the cell is a breast tissue derived cell, and the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:36, 44, 45, 89, 146, 154, 159, 165, 172, 174, 183, 203, 261, 364, 366, 387, 419, 420, 496, 503, 510, 512, 529, 552, 560,564, 570, 590, 606, 644, 646, 693, 707, 711, 726, 746, 754, 756, 875, 902, 921, 942, 990, 1095, 1104, 1122, 1131, 1142, 1170, 1184, 1205, 1286, 1289, 1354, 1387, 1435, 1535, 1751, 1764, 1777, 1795, 1860, 1869, 1882, 1890, 1915, 1933, 1934, 1979, 1980, 2007, 2023, 2040, 2059, 2223, 2245, 2300, 2325, 2409, 2462, 2486, 2488, and 2492.

19. The method of claim 16, wherein the cell is a colon tissue derived cell, and the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:33, 65, 228, 250, 252, 253, 280, 282, 355, 370, 387, 443, 460, 491, 545, 560, 581, 603, 680, 693, 703, 704, 716, 726, 746, 752, 753, 1095, 1104, 1205, 1241, 1264, 1354, 1387, 1401, 1442, 1514, 1734, 1742, 1780, 1851, 1899, 1915, 1954, 2024, 2066, 2262, and 2325.

20. The method of claim 16, wherein the cell is a lung tissue derived cell, and the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS: 10, 54, 65, 171, 174, 203, 252, 253, 254, 285, 419, 420, 466, 491, 525, 526, 552, 571, 574, 590, 693, 700, 726, 742, 746, 861, 922, 990, 1088, 1288, 1355, 1417, 1422, 1444, 1454, 1570, 1597, 1979, 2007, 2024, 2034, 2038, 2126, and 2245.

21. The method of claim 16, wherein the differentially expressed gene product is encoded by a gene corresponding to a sequence of at least one of SEQ ID NOS:648 and 1899.

<212> DNA
<213> Homo sapiens

<400> 1075
ggcaccccca agatgttttc ttcttaatta ttcctaaata cttttatgtg ttggcattaa 60
attgtaactt tataggctcc cctattcttt ttgctttttt ttccccctga aattactgag 120
caacaagatt cctgttctct ccccttcaag gctttgtttt ctggaacttg acattctcaa 180
atcattgcca gttatttta gtacgtgatt agtctccctt cctcaggtat gttttcccca 240
atttggattg aatctactgt ttgcatcttg tttcccatcc caccttcata cagattgtat 300

<210> 1076
<211> 300
<212> DNA
<213> Homo sapiens

<400> 1076
tgctaattca gccctaaacc ccatcctcta caacatgaca ctgtgcagga atgagtggaa 60
gaaaattttt tgctgcttct ggttcccaga aaagggagcc attttaacag acacatctgt 120
caaaagaaat gacttgtcga ttatttctgg ctaatttttc tttatagccg agtttctcac 180
acctggcgag ctgtggcatg cttttaaaca gagttcattt ccagtaccct ccatcagtgc 240
accctgcttt aagaaaatga acctatgcaa atagacatcc acagcgtcgg taaattaagg 300

<210> 1077
<211> 300
<212> DNA
<213> Homo sapiens

<400> 1077
taagtgggct aagaccagaa gagagactta ttcgcttaag tagaaacatg tgccttttat 60
taactgcagt cctgcatttt atccatggaa tgacagaccc tgtattaatg tctctcagtg 120
cctctcatgt gtcatctttt cgtagacatt ttcctgtgct gtttgtctct gcttgcctgt 180
ttattcttcc tgtcttactc agttatgttc tttggcatca ctatgcacta aatacatggt 240
tgtttgcagt tacagcattt tgtgtggaac tgtgcttaaa agtaattgtt tctctcactg 300

<210> 1078
<211> 300
<212> DNA
<213> Homo sapiens

<400> 1078
gtcagatgtt tctggggacg ttgagctgca gtgaagtgag aggggcagag ggggcttttg 60
aagtcacaag gtcagggaga ggagaagaag cgtgctggat gagtcacact gtaggactca 120
agccagtagg ttcttgttag cccggctact gacctggagc caggcactga tagcaacgtg 180
tcctctgagg gaaggcaaat gggaaatcca agcaggcact gggatctgcc tgtgacactc 240
ttgtggggcc tggtccctcg acctaagtga gcttgggcca ctcagagcca ccccaggtgc 300

<210> 1079
<211> 300
<212> DNA
<213> Homo sapiens

<400> 1079
gggcgaagaa ggctggttgg gaaggagacc agcataaaca ctttggggac tgagaggata 60
agccatatca ttagtgaccc tcggcagaaa gaaaagaata aagcgttggc ttctgatttt 120
cctcacattt ctgcttgtgc acatgagaca ggcaaatgta cactggggac caccatgttc 180
acgtgacatc aagaggaagc ggaaaccagt ggccacagca tctttgtcta gccccagtgc 240

This Page Blank (uspto)